MOLECULAR BIOLOGY OF
THE CELL
THIRD EDITION

Bruce Alberts • Dennis Bray
Julian Lewis • Martin Raff • Keith Roberts
James D. Watson

Garland Publishing, Inc.
New York & London

GARLAND STAFF
Text Editor: Miranda Robertson
Managing Editor: Ruth Adams
Illustrator: Nigel Orme
Molecular Model Drawings: Kate Hesketh-Moore
Director of Electronic Publishing: John M-Roblin
Computer Specialist: Chuck Bartelt
Disk Preparation: Carol Winter
Copy Editor: Shirley M. Cobert
Production Editor: Douglas Goertzen
Production Coordinator: Perry Bessas
Indexer: Maija Hinkle

Bruce Alberts received his Ph.D. from Harvard University and is currently President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. *Dennis Bray* received his Ph.D. from the Massachusetts Institute of Technology and is currently a Medical Research Council Fellow in the Department of Zoology, University of Cambridge. *Julian Lewis* received his D.Phil. from the University of Oxford and is currently a Senior Scientist in the Imperial Cancer Research Fund Developmental Biology Unit, University of Oxford. *Martin Raff* received his M.D. from McGill University and is currently a Professor in the MRC Laboratory for Molecular Cell Biology and the Biology Department, University College London. *Keith Roberts* received his Ph.D. from the University of Cambridge and is currently Head of the Department of Cell Biology, the John Innes Institute, Norwich. *James D. Watson* received his Ph.D. from Indiana University and is currently Director of the Cold Spring Harbor Laboratory. He is the author of *Molecular Biology of the Gene* and, with Francis Crick and Maurice Wilkins, won the Nobel Prize in Medicine and Physiology in 1962.

Library of Congress Cataloging-in-Publication Data
Molecular biology of the cell / Bruce Alberts . . . [et al.].—3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-8153-1619-4 (hard cover).—ISBN 0-8153-1620-8 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology. QH 581.2 M718 1994]
QH581.2.M64 1994
574.87—dc20
DNLM/DLC
for Library of Congress
93-45907
CIP

Published by Garland Publishing, Inc.
717 Fifth Avenue, New York, NY 10022

Printed in the United States of America

15 14 13 12 10 9 8 7 6 5 4 3 2

Front cover: The photograph shows a rat nerve cell in culture. It is labeled (*yellow*) with a fluorescent antibody that stains its cell body and dendritic processes. Nerve terminals (*green*) from other neurons (not visible), which have made synapses on the cell, are labeled with a different antibody. (Courtesy of Olaf Mundigl and Pietro de Camilli.)

Dedication page: Gavin Borden, late president of Garland Publishing, weathered in during his mid-1980s climb near Mount McKinley with MBoC author Bruce Alberts and famous mountaineer guide Mugs Stump (1940–1992).

Back cover: The authors, in alphabetical order, crossing Abbey Road in London on their way to lunch. Much of this third edition was written in a house just around the corner. (Photograph by Richard Olivier.)


carrot

extracts. If these minor cell proteins differ among cells to the same extent as the more abundant proteins, as is commonly assumed, only a small number of protein differences (perhaps several hundred) suffice to create very large differences in cell morphology and behavior.

A Cell Can Change the Expression of Its Genes in Response to External Signals [3]

Most of the specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of several specific proteins is dramatically increased. Glucocorticoids are released during periods of starvation or intense exercise and signal the liver to increase the production of glucose from amino acids and other small molecules; the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids in different ways. In fat cells, for example, the production of tyrosine aminotransferase is reduced, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization—different cell types often respond in different ways to the same extracellular signal. Underlying this specialization are features that do not change, which give each cell type its permanently distinctive character. These features reflect the persistent expression of different sets of genes.

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein [4]

The differences between the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? There are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the primary RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytoplasm (**RNA transport control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 9–2).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 9–2, only transcriptional control ensures that no superfluous intermediates are synthesized. In the

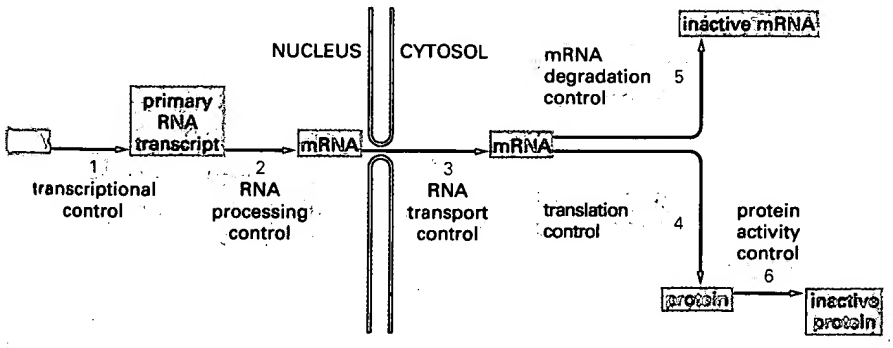


Figure 9–2 Six steps at which eucaryote gene expression can be controlled. Only controls that operate at steps 1 through 5 are discussed in this chapter. The regulation of protein activity (step 6) is discussed in Chapter 5; this includes reversible activation or inactivation by protein phosphorylation as well as irreversible inactivation by proteolytic degradation.

following sections we discuss the DNA and protein components that regulate the initiation of gene transcription. We return at the end of the chapter to the other ways of regulating gene expression.

Summary

The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.

DNA-binding Motifs in Gene Regulatory Proteins [5]

How does a cell determine which of its thousands of genes to transcribe? As discussed in Chapter 8, the transcription of each gene is controlled by a regulatory region of DNA near the site where transcription begins. Some regulatory regions are simple and act as switches that are thrown by a single signal. Other regulatory regions are complex and act as tiny microprocessors, responding to a variety of signals that they interpret and integrate to switch the neighboring gene on or off. Whether complex or simple, these switching devices consist of two fundamental types of components: (1) short stretches of DNA of defined sequence and (2) *gene regulatory proteins* that recognize and bind to them.

We begin our discussion of gene regulatory proteins by describing how these proteins were discovered.

Gene Regulatory Proteins Were Discovered Using Bacterial Genetics [6]

Genetic analyses in bacteria carried out in the 1950s provided the first evidence of the existence of **gene regulatory proteins** that turn specific sets of genes on or off. One of these regulators, the *lambda repressor,* is encoded by a bacterial virus, *bacteriophage lambda.* The repressor shuts off the viral genes that code for the protein components of new virus particles and thereby enables the viral genome to remain a silent passenger in the bacterial chromosome, multiplying with the bacterium when conditions are favorable for bacterial growth (see Figure 6–80). The lambda repressor was among the first gene regulatory proteins to be characterized, and it remains one of the best understood, as we discuss later. Other bacterial regulators respond to nutritional conditions by shutting off genes encoding specific sets of metabolic enzymes when they are not needed. The *lac repressor,* for example, the first of these bacterial proteins to be recognized, turns off the production of the proteins responsible for lactose metabolism when this sugar is absent from the medium.

The first step toward understanding gene regulation was the isolation of mutant strains of bacteria and bacteriophage lambda that were unable to shut off specific sets of genes. It was proposed at the time, and later proved, that most of these mutants were deficient in proteins acting as specific repressors for these sets of genes. Because these proteins, like most gene regulatory proteins, are present in small quantities, it was difficult and time-consuming to isolate them. They were eventually purified by fractionating cell extracts on a series of standard chromatography columns (see pp. 166–169). Once isolated, the proteins were shown to bind to specific DNA sequences close to the genes that they

Figure 9–3 **Double-helical structure of DNA.** The major and minor grooves on the outside of the double helix are indicated. The atoms are colored as follows: carbon, *dark blue* or *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.*

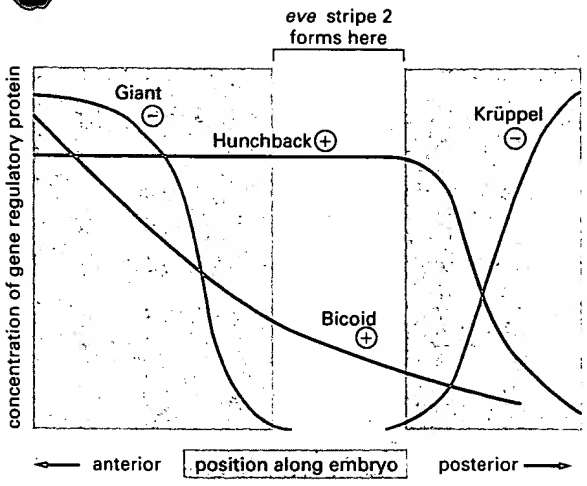


Figure 9–43 Distribution of the gene regulat ry proteins responsible for ensuring that *eve* is expressed in stripe 2. The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins (Figure 9–39). The expression of *eve* in stripe 2 occurs only at the position where the two activators (Bicoid and Hunchback) are present and the two repressors (Giant and Krüppel) are absent. In fly embryos that lack Krüppel, for example, stripe 2 expands posteriorly. Likewise, stripe 2 expands posteriorly if the DNA-binding sites for Krüppel in the stripe 2 module (see Figure 9–41) are inactivated by mutation and this regulatory region is reintroduced into the genome.

The *eve* gene itself encodes a gene regulatory protein, which, after its pattern of expression is set up in seven stripes, in turn regulates the expression of other *Drosophila* genes. As development proceeds, the embryo is thus subdivided into finer and finer regions that eventually give rise to the different body parts of the adult fly, as discussed in Chapter 21.

strung out over 20,000 nucleotide pairs of DNA and binds more than 20 different proteins. A large and complex control region is thereby built from a series of smaller modules, each of which consists of a unique arrangement of short DNA sequences recognized by specific gene regulatory proteins. An important requirement of this strategy is the absence of cross-talk between the modules: the state of one module should not affect that of the others. How this molecular insulation is achieved is unknown. In this way, however, a single gene can respond to an enormous number of combinatorial inputs.

Complex Mammalian Gene Control Regions Are Also Constructed from Simple Regulatory Modules [33]

It has been estimated that several percent of the coding capacity of a mammalian genome is devoted to the synthesis of proteins that serve as regulators of gene transcription. This reflects the exceedingly complex controls that regulate the expression of mammalian genes. It is not unusual, for example, to find a gene with a control region that is 50,000 nucleotide pairs in length, in which many modules, each containing a number of regulatory sequences that bind gene regulatory proteins, are interspersed with long stretches of spacer DNA.

One of the best-understood examples of a complex mammalian regulatory region is found in the human β-globin gene, which is expressed exclusively in red blood cells and at a specific time in their development. A complex array of gene regulatory proteins controls the expression of the gene, some acting as activators and others as repressors (Figure 9–45). The concentrations (or activities) of many of these gene regulatory proteins are thought to change during development, and

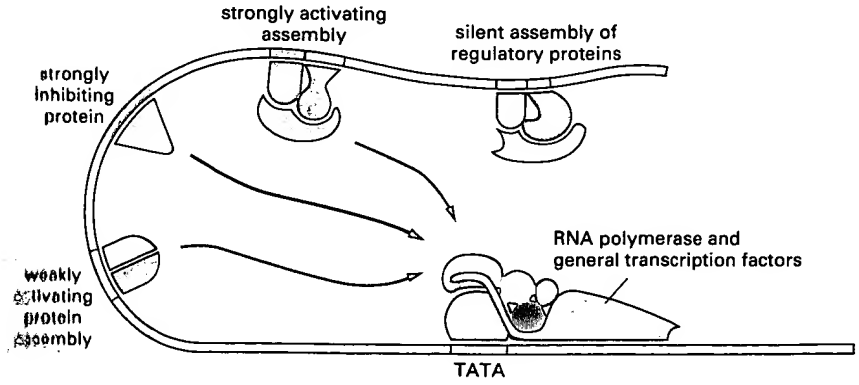


Figure 9–44 Integration at a promoter. Multiple sets of gene regulatory proteins can work together to influence a promoter, as they do in the *eve* stripe 2 module illustrated previously in Figure 9–42. It is not yet understood in detail how the integration of multiple inputs is achieved.

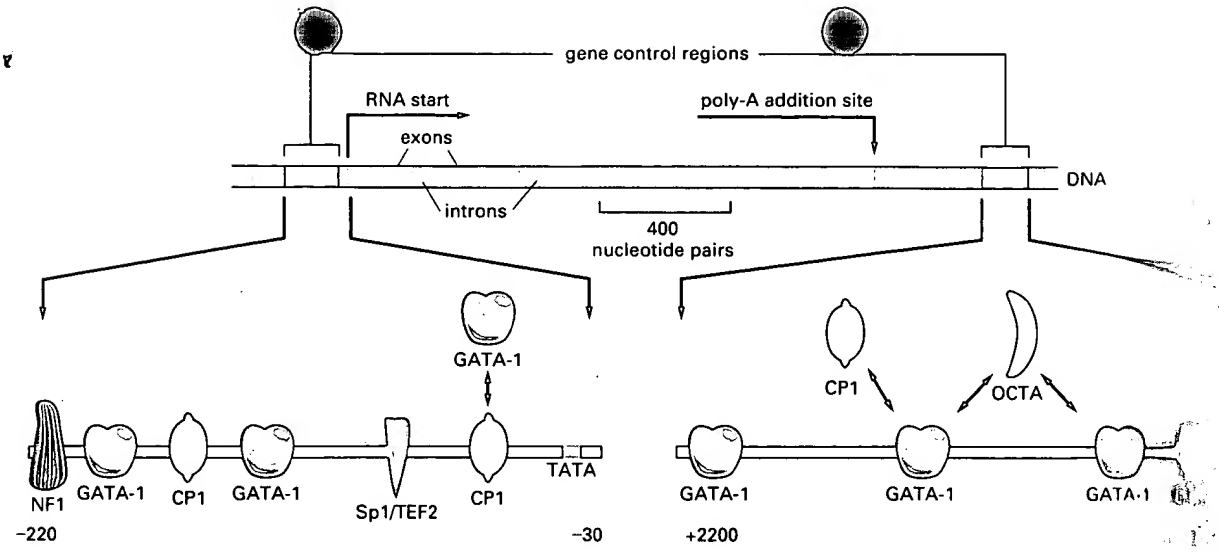


Figure 9–45 **Model for the control of the human β-globin gene.** This diagram shows some of the gene regulatory proteins thought to regulate expression of the gene during red blood cell development (see Figure 9–52). Some of the gene regulatory proteins shown, such as CP1, are found in many types of cells, while others, such as GATA-1, are present in only a few types of cells including red blood cells and therefore are thought to contribute to the cell-type specificity of β-globin gene expression. As indicated by the double-headed arrows, several of the binding sites for GATA-1 overlap those of other gene regulatory proteins; it is thought that occupancy of these sites by GATA-1 excludes binding of other proteins. (Adapted from B. Emerson, In Gene Expression: General and Cell-Type Specific (M. Karin, ed.), pp. 116–161. Boston: Birkhauser, 1993.)

only a particular combination of all the proteins triggers transcription of the gene. We see later that the β-globin gene is also subject to a second, higher layer of control that involves global changes in chromatin structure.

The Activity of a Gene Regulatory Protein Can Itself Be Regulated [34]

The strategies for regulating the *eve* gene and the human β-globin gene are similar in that the gene control regions respond to a bewildering array of gene regulatory proteins. *Drosophila* is unusual, however, in the way that the spatial distribution of gene regulatory proteins in the cytoplasm controls gene expression. As discussed previously, the early *Drosophila* embryo is a single giant cell that contains thousands of nuclei in a common cytoplasm, and the gene regulatory proteins themselves are distributed in complex spatial patterns so that different nuclei are exposed to different concentrations of the proteins. These gene regulatory proteins enter the nuclei directly to activate or repress transcription of their target genes. In most embryos of other organisms, individual nuclei are in separate cells, and extracellular positional information must either pass across the plasma membrane or, more usually, generate signals in the cytosol in order to influence the genome.

The mechanisms by which extracellular signals communicate their message across the plasma membrane to gene regulatory proteins inside the cell are discussed in Chapter 15. Here we need deal only with the final steps in the intracellular signaling cascades activated by extracellular signals—the steps in which the activity of gene regulatory proteins is altered. In many cases the gene regulatory protein is present in the cell in an inactive form and a signal alters the protein so as to activate it. The protein may be activated by phosphorylation catalyzed by a protein kinase, for example, or it may be released from a tight complex with a second protein that otherwise holds the gene regulatory protein in the cytosol, preventing it from entering the nucleus. These and some other ways of controlling the activity of gene regulatory proteins are illustrated in Figure 9–46.

Bacteria Use Interchangeable RNA Polymerase Subunits to Help Regulate Gene Transcription [35]

We have seen the importance of gene regulatory proteins that bind to regulatory sequences in DNA and signal to the transcription apparatus whether or not to start the synthesis of an RNA chain. Although this is the main way of controlling transcriptional initiation in both eucaryotes and procaryotes, some bacteria and

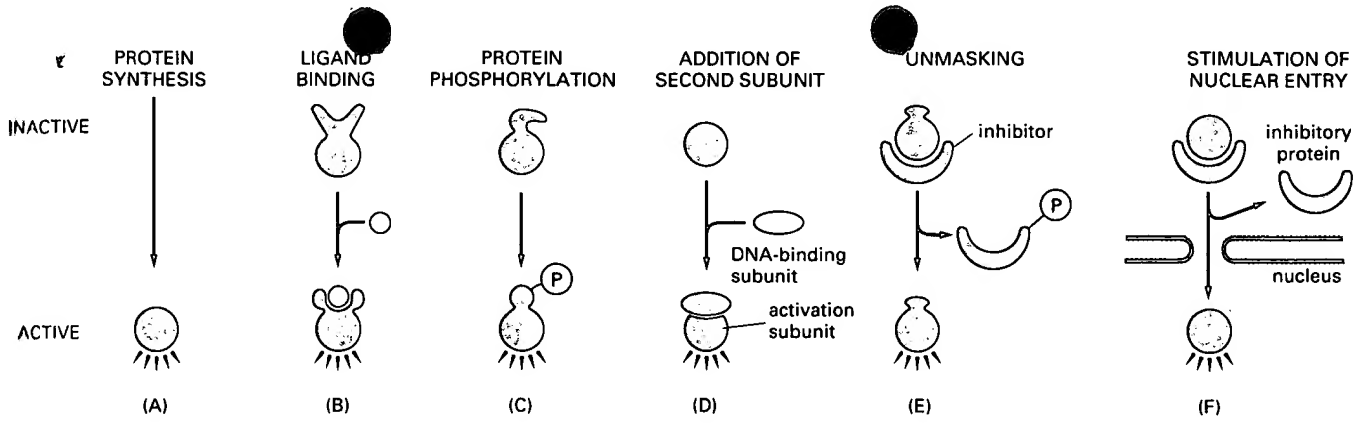


Figure 9–46 **Some ways in which the activity of gene regulatory proteins is regulated in eucaryotic cells.** (A) The protein is synthesized only when needed and is rapidly degraded by proteolysis so that it does not accumulate. (B) Activation by ligand binding. (C) Activation by phosphorylation. (D) Formation of a complex between a DNA-binding protein and a separate protein with a transcription-activating domain. (E) Unmasking of an activation domain by the phosphorylation of an inhibitor protein. (F) Stimulation of nuclear entry by removal of an inhibitory protein that otherwise keeps the regulatory protein from entering the nucleus.

their viruses use an additional strategy based on interchangeable subunits of RNA polymerase. As described in Chapter 8, a *sigma (σ) subunit* is required for the bacterial RNA polymerase to recognize a promoter. Some bacteria make several different sigma subunits, each of which can interact with the RNA polymerase core and direct it to different specific promoters. This scheme permits one large set of genes to be turned off and a new set to be turned on simply by replacing one sigma subunit with another. The strategy is efficient because it bypasses the need to deal with the genes one by one, and it is often used by bacterial viruses to activate several sets of genes rapidly and sequentially (Figure 9–47).

In a sense, eucaryotes employ an analogous strategy through the use of three distinct RNA polymerases (I, II, and III) that share some of their subunits. Procaryotes, in contrast, use only one type of core RNA polymerase molecule but modify it with different sigma subunits.

Gene Switches Have Gradually Evolved

We have seen that the control regions of eucaryotic genes are often spread out over long stretches of DNA, whereas those of procaryotic genes are typically closely packed around the start point of transcription. Several bacterial gene regulatory proteins, however, recognize DNA sequences that are located many nucleotide pairs away from the promoter. The example of DNA looping in *E. coli* shown previously in Figure 9–32 resembles the way that eucaryotic gene regulatory proteins act at a distance. In fact, this case provided one of the first examples of DNA looping in gene regulation and greatly influenced later studies of eucaryotic gene regulatory proteins.

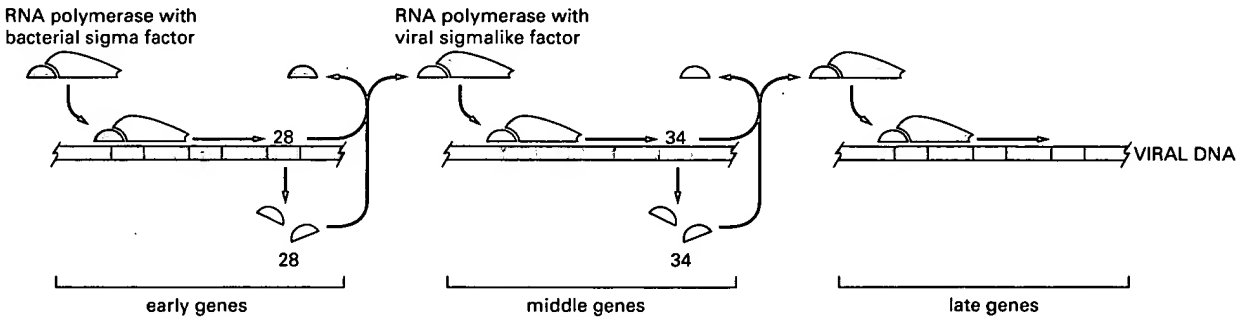


Figure 9–47 **Interchangeable RNA polymerase subunits as a strategy to control gene expression in a bacterial virus.** The bacterial virus SPO1, which infects the bacterium *B. subtilis*, uses the bacterial polymerase to transcribe its early genes. One of the early genes, called 28, encodes a sigmalike factor that binds to RNA polymerase and displaces the bacterial sigma factor. This new form of polymerase specifically initiates transcription of the SPO1 "middle" genes. One of the middle genes encodes a second sigmalike factor that displaces the 28 product and directs RNA polymerase to transcribe the "late" genes. This last set of genes produces the proteins that package the virus chromosome into a virus coat and lyse the cell. Thus, by this strategy, sets of virus genes are expressed in a particular order, allowing for rapid, yet temporally controlled, viral replication.

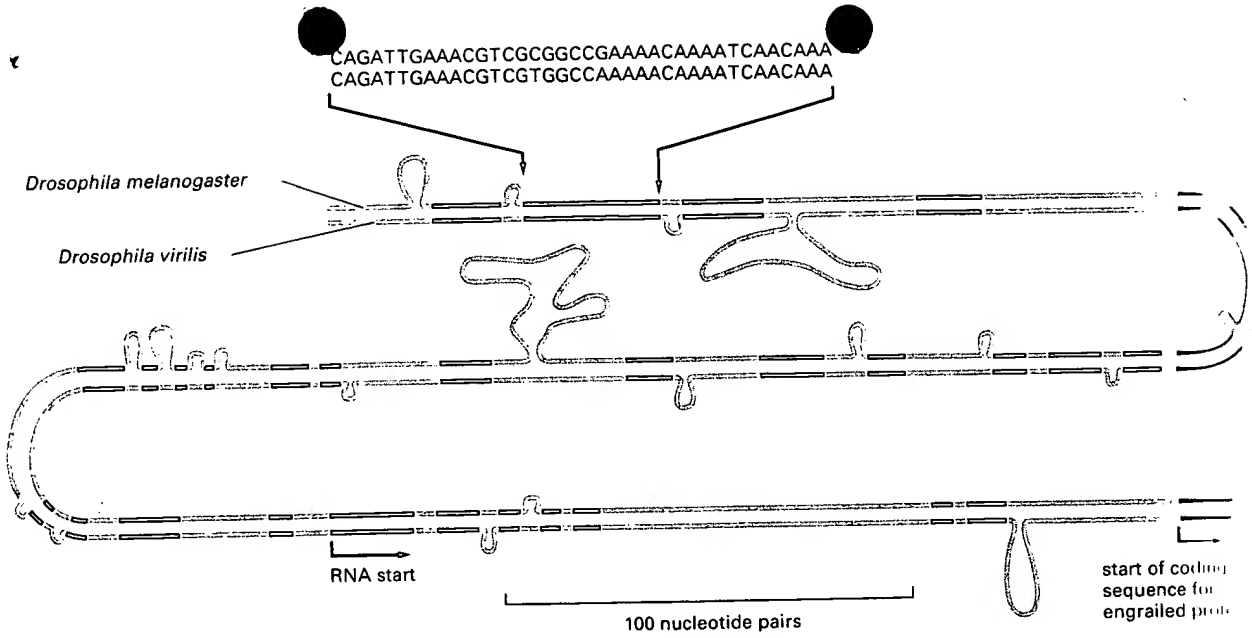


Figure 9–48 **A comparison of part of the control region upstream from the *engrailed* gene in two species of *Drosophila*.** A DNA sequence comparison between *Drosophila melanogaster* and *Drosophila virilis* is shown, with regions of 90% sequence conservation shown in *red*. One example of an actual sequence match is illustrated in detail at the top. The conserved sequences presumably mark the sites where important gene regulatory proteins bind, whereas loops indicate places where insertions or deletions of nucleotides have occurred since these two species evolved from a common ancestor about 60 million years ago. (Courtesy of Judith A. Kassis and Patrick H. O'Farrell.)

It seems likely that the close-packed arrangement of bacterial genetic switches developed from more extended forms of switches in response to the evolutionary pressure on bacteria to maintain a small genome size. (The same argument has been used to explain the lack of introns in bacteria, as discussed in Chapter 8.) This compression comes at a price, however, as it is difficult to imagine how the compact switches could be easily altered to incorporate new levels of control. The extended form of eucaryotic control regions, in contrast, with discrete regulatory modules separated by long stretches of spacer DNA, would be expected to facilitate reshuffling of modules during evolution, both to create new regulatory circuits and to modify old ones. Unraveling the history of how gene control regions evolved presents a fascinating challenge, and many clues can be found in present-day DNA sequences (Figure 9–48).

Summary

The transcription of individual genes is switched on and off in cells by gene regulatory proteins. In procaryotes these proteins usually bind to specific DNA sequences close to the RNA polymerase start site and, depending on the nature of the regulatory protein and the precise location of its binding site relative to the start site, either activate or repress transcription of the gene. The flexibility of the DNA helix, however, also allows proteins bound at distant sites to affect the RNA polymerase at the promoter by the looping out of the intervening DNA. Such action at a distance is extremely common in eucaryotic cells, where gene regulatory proteins bound to sequences thousands of nucleotide pairs from the promoter can control gene expression.

Although procaryotic RNA polymerases can initiate transcription on their own, eucaryotic polymerases require the prior assembly of general transcription factors at the promoter. These factors assemble in a particular order, beginning with the binding of TFIID to the TATA box, a DNA sequence found just upstream of most eucaryotic RNA polymerase start sites. The ordered assembly of general transcription factors provides several steps at which the initiation of transcription can be regulated, and many eucaryotic gene regulatory proteins are thought to work by facilitating (positive control) or hindering (negative control) the assembly process.

Whereas the transcription of a typical procaryotic gene is controlled by only one or two gene regulatory proteins, the regulation of higher eucaryotic genes is much more complex, commensurate with the larger genome size and the large variety of cell types. The control region of the Drosophila eve *gene, for example, encompasses 20,000 nucleotide pairs of DNA and has binding sites for over 20 gene regulatory proteins.*

Some of these proteins are transcriptional activators, while others are transcriptional repressors. These proteins bind to regulatory sequences organized in a series of regulatory modules strung together along the DNA.

Chromatin Structure and the Control of Gene Expression [36]

As discussed in Chapter 8, the genomes of eucaryotes are highly compacted to allow the very long DNA molecules to fit inside the cell and to be managed easily. The first level of compaction is the wrapping of DNA around histones to form nucleosomes. In a second level of compaction nucleosomes are packed into 30-nm filaments. Finally, an even higher order of packing (still poorly understood) is observed in heterochromatin, which is confined to selected regions of the genome that show an unusually condensed interphase structure.

How do gene regulatory proteins and the general transcription factors gain access to DNA that is packed into these compact protein-DNA structures, and how does the packing affect the control of gene expression? We see in this section that two general principles have emerged from studies of chromatin structure and its influence on gene expression. First, nucleosomes do not usually present a serious obstruction to either gene regulatory proteins or RNA polymerases. Enhancers can still function despite them, histones that block a promoter can be displaced, and, once transcription has begun, Pol II can transcribe through the nucleosomes without dislodging them. Even bacterial polymerases, which do not encounter nucleosomes *in vivo*, can transcribe through them, suggesting that the nucleosome is built to be traversed easily (Figure 9–49). The second general principle is that some forms of higher-order DNA packaging render the DNA inaccessible both to gene regulatory proteins and to the general transcription factors. Higher-order DNA packaging thus plays a crucial part in the control of gene expression in eucaryotes, serving to silence large sections of the genome—in some cases reversibly, in other cases not.

Transcription Can Be Activated on DNA That Is Packaged into Nucleosomes [37]

In the previous section we described a simple model for how transcription of a eucaryotic gene is activated by a gene regulatory protein (see Figure 9–36). How

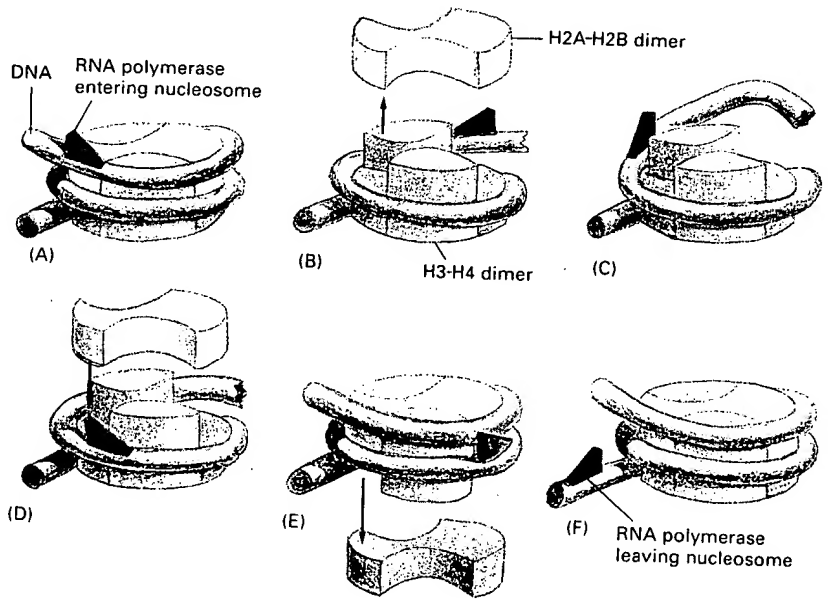


Figure 9–49 **A tentative model to account for the ability of RNA polymerases to transcribe through nucleosomes without causing their displacement.** The polymerase first displaces an H2A-H2B dimer (see Figure 8–10), allowing the polymerase to enter the nucleosome. In the next step the polymerase pulls the DNA away from the H3-H4 dimer it next encounters and continues transcribing. The displaced H2A-H2B dimer is recaptured by the nucleosome, and the second, symmetrically disposed H2A-H2B dimer is now displaced, allowing the process to repeat and permitting the polymerase to exit from the nucleosome. (After K.E. van Holde et al., *J. Biol. Chem.* 267:2837–2840, 1992.)

must this model be modified to take account of the presence of nucleosomes? To begin with the first step, how do nucleosomes affect the binding of a gene activator protein to its regulatory DNA sequence? In some cases the regulatory sequences reside in short nucleosome-free regions, and so no problem arises. It is uncertain how such nucleosome-free regions are maintained, but, as discussed in Chapter 8, some stretches of DNA are too stiff to accommodate the tight folding necessary for nucleosome formation. In other cases, however, the regulatory sequence is packaged into a nucleosome and yet at least some gene activator proteins can still recognize and bind to it. Once bound, the regulatory proteins appear to destabilize the nucleosome, which is then at least partially disassembled. Which types of gene regulatory proteins can achieve this feat and how they accomplish it remain unknown.

The general transcription factors, in contrast, seem unable to assemble onto a promoter that is packaged into a nucleosome. In fact, such packaging may have evolved in part to ensure that leaky, or basal, transcription initiation (that is, initiation without a gene activator protein bound upstream) does not occur. The binding of a gene activator protein thousands of nucleotide pairs away from a nucleosome-packaged promoter, however, can apparently displace a nucleosome from a promoter and thereby allow the assembly of the general transcription factors. The displacement either could be due to a separate activity of the gene activator protein or could be an indirect consequence of the activator contacting the general transcription factors to facilitate their assembly on the DNA (Figure 9–50).

Some Forms of Chromatin Silence Transcription [38]

Although transcription can occur on DNA that is packaged into nucleosomes, the DNA in some special forms of chromatin appears to be inaccessible to gene ac-

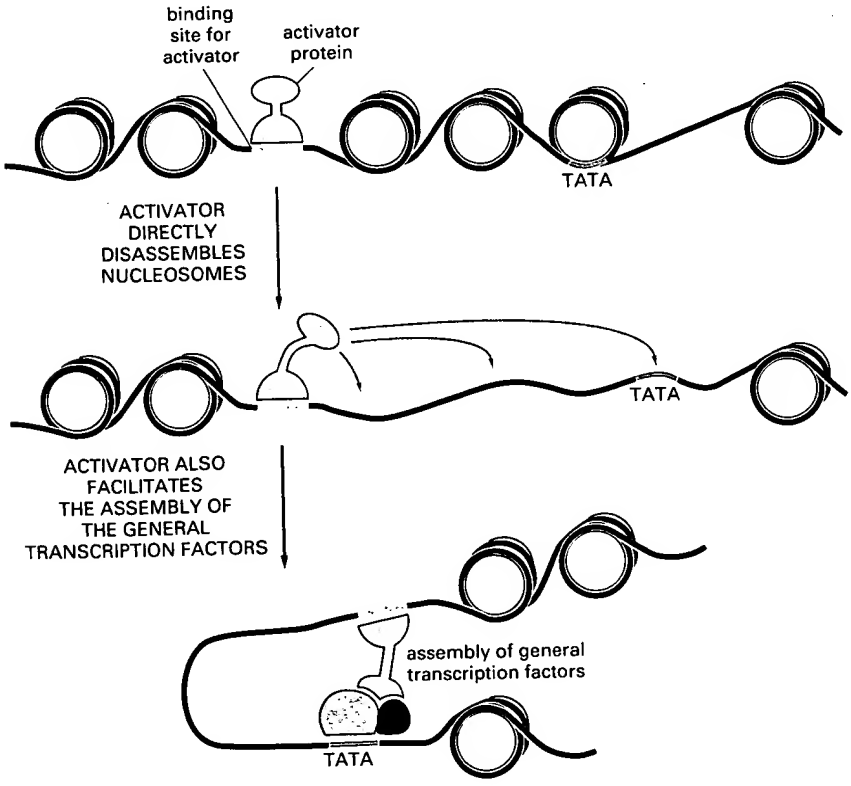


Figure 9–50 **One model to explain the displacement of nucleosomes during the initiation of transcription in eucaryotes.** A bound gene activator protein possesses a separate activity that directly removes nucleosomes from the promoter, exposing the promoter to the general transcription factors and enabling them to assemble. In a different model (not shown) nucleosome displacement occurs as an indirect consequence of the activator protein promoting the assembly of the general transcription factors; the activator protein, for example, could permit the general factors to begin to assemble in the presence of a nucleosome, and, once partially assembled, these factors would destabilize the nucleosome. The process underlying nucleosome displacement is poorly understood. The histones may simply leave the DNA or, according to an alternative model, they may disassemble but remain bound to DNA (see Figure 8–40).

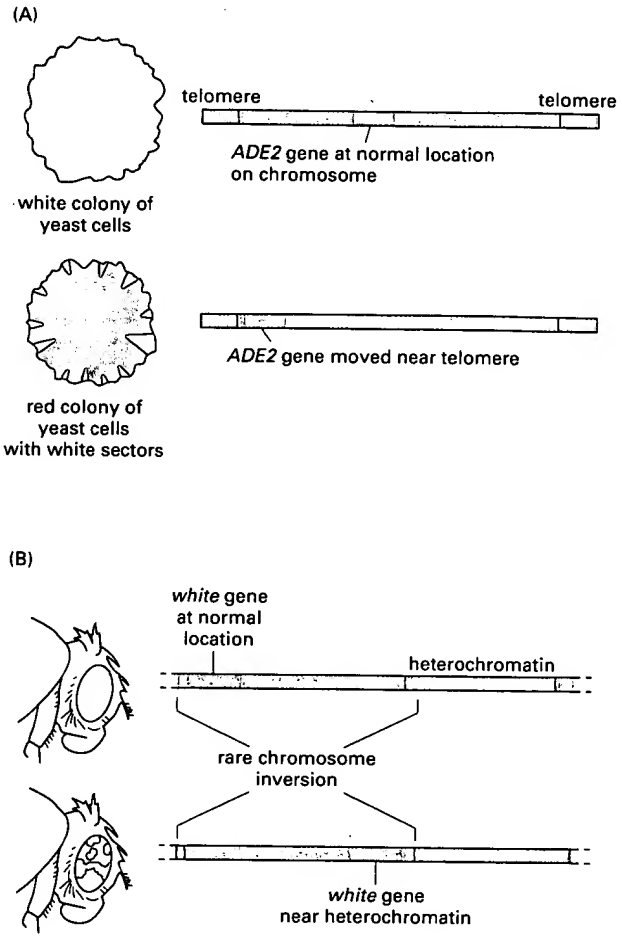


Figure 9–51 Position effects on gene expression. (A) The yeast *ADE2* gene at its normal chromosomal location is expressed in all cells. When moved near the end of a yeast chromosome, the gene is silenced in most but not all cells of the population. The absence of the *ADE2* gene product results in a block in the adenine biosynthetic pathway, which leads to the accumulation of a red pigment. The founder cell for the sectored colony shown had its *ADE2* gene shut off, so most of the colony is *red*. Normally, yeast colonies are white. The *white* sectors at the edges of the *red* colony are clones of cells where the *ADE2* gene has spontaneously become active. The finding of such sectoring indicates that the active and inactive states of *ADE2* expression are heritable when this gene is near the telomere, a topic discussed in more detail in the next section.

(B) Position effects can also be observed for the *Drosophila white* gene. Wild-type flies with a normal *white* gene have red eyes. If the *white* gene is inactivated by mutation, the eyes become white (hence the name of the gene). In flies with a chromosomal inversion that moves the *white* gene near a heterochromatic region, the eyes are mottled, with red and white patches. The white patches represent cells where the *white* gene is silenced and red patches represent cells that express the *white* gene. The difference is thought to arise from variations in how far along the chromosome the heterochromatin spreads early in eye development. As in the case of yeast *ADE2* gene, once established, the state of *white* expression is heritable, producing patches of many cells that express *white* as well as patches of cells where *white* is silenced. (After L.L. Sandell and V.A. Zakian, *Trends Cell Biol.* 2:10–14, 1992.)

...tor proteins. These *inactive* forms of chromatin, including the especially ...bly condensed form called *heterochromatin* (discussed in Chapter 8), are as... ...ned to contain special proteins that make the DNA unusually inaccessible.

An observation in the yeast *S. cerevisiae* illustrates how some types of chro... ...tin can shut off gene transcription. The *ADE2* gene, whose expression is par... ...ularly easy to monitor, is expressed when present at its normal chromosomal ...ation. When this gene is experimentally relocated to the end of a chromosome, ...wever, its transcription is turned off, even though the cell contains all of the ...teins required to transcribe the gene. The DNA near the ends of yeast chro... ...somes (the *telomeres*) is packaged into an especially inaccessible form of ...omatin, and it is this packaging that is thought to be responsible for maintain... ...the translocated *ADE2* gene in an inactive state, a process called *silencing*. The ...encing of genes located near chromosome ends extends for approximately ...000 nucleotide pairs and applies to many genes in addition to *ADE2*; the ...encing seems to weaken gradually with distance from the telomere. The ...hanism of silencing is not known, but it seems likely to involve a coopera... ...assembly of proteins on the DNA that, once established, is heritable following ...A replication (Figure 9–51A).

The silencing of the *ADE2* gene is an example of a **position effect,** in which ...activity of a gene is dependent on its position in the genome. Position effects ...re first recognized in *Drosophila* (Figure 9–51B), but they have now been ob... ...ved in a number of other organisms and are thought to reflect the different ...tes of chromatin present at different locations in the genome and the tendency ...these states to spread to encompass nearby genes. We revisit this topic later ...the chapter when we analyze mechanisms of cell memory.

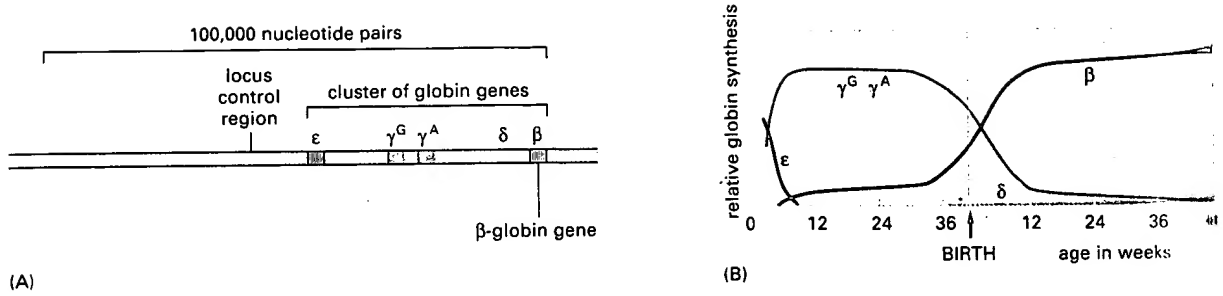


Figure 9–52 **The cluster of β-like globin genes in humans.** (A) The large chromosomal region shown spans 100,000 nucleotide pairs and contains the five globin genes and a locus control region (discussed in the text). (B) Changes in the expression of the β-like globin genes at various stages of human development. Each of the globin chains encoded by these genes combines with an α-globin chain to form the hemoglobin in red blood cells. (A, after F. Grosveld, G.B van Assendelft, D.R. Greaves, and G. Kollias, *Cell* 51:975–985, 1987. © Cell Press.)

An Initial Decondensation Step May Be Required Before Mammalian Globin Genes Can Be Transcribed [39]

Another example of how highly condensed chromatin can prevent gene expression comes from studies of chick and human β-globin gene clusters. The five genes of the cluster, spread over 50,000 nucleotide pairs of DNA, are transcribed exclusively in erythroid cells (that is, cells of the red blood cell lineage). Moreover, each gene is turned on at a different stage of development (Figure 9–52) and in different organs: the ε-globin gene is expressed in the embryonic yolk sac, γ in the yolk sac and the fetal liver, and δ and β primarily in the adult bone marrow. We previously described a series of gene regulatory proteins that are necessary to turn on the human β-globin gene at the appropriate time and place (see Figure 9–45), and each of the other globin genes has a similar set of regulatory proteins, many of which are shared among these genes. In addition to the individual regulation of each of the globin genes, however, the entire cluster appears to be subject to an on-off control that involves global changes in chromatin structure.

Some of the first evidence for such changes came from studies of the sensitivity of the globin genes in isolated nuclei to digestion by the nuclease enzyme DNaseI. In cells where the globin genes are not expressed, the DNA in these genes is resistant to DNaseI, indicating that they are tightly packaged into chromatin. In erythroid cells, by contrast, the entire gene cluster is sensitive to DNaseI, indicating that the chromatin has changed to make the DNA more accessible to the enzyme. The DNA is still folded into nucleosomes, but the higher-order packing of the chromatin has loosened. This change in DNA packing occurs even before the individual globin genes are transcribed, suggesting that the genes are regulated in two steps. In the first step the chromatin of the entire globin locus is decondensed, which is presumed to allow some of the gene regulatory proteins access to the DNA. In the second step the remaining gene regulatory proteins assemble on the DNA and direct the expression of individual genes (Figure 9–53).

The extensive change in chromatin structure that occurs in the first step is thought to require a region of DNA (called the *locus control region,* or *LCR*) that lies far upstream from the gene clusters (see Figure 9–52). The importance of the LCR can be seen in patients with a certain type of thalassemia, a severe genetic form of anemia. In these patients the β-globin locus is found to have undergone deletions that remove all or part of the LCR, and although the β-globin gene and its nearby regulatory regions are intact, the gene remains silent in erythroid cells. Moreover, the β-globin gene in the erythroid cells remains DNaseI resistant, indicating that it fails to undergo the normal chromatin decondensation step during erythroid cell development.

Subsequent experiments in transgenic mice have confirmed the profound effects of the LCR on the expression of globin genes. When, for example, the human β-globin gene plus its local regulatory sequences (the region shown in Figure 9–45) is inserted into different positions in the mouse genome, it is expressed at low levels that depend on the site of insertion. This behavior is typical for mammalian genes, and it indicates that local position effects influence the expression of the gene (Table 9–2). When the LCR is included with the gene,

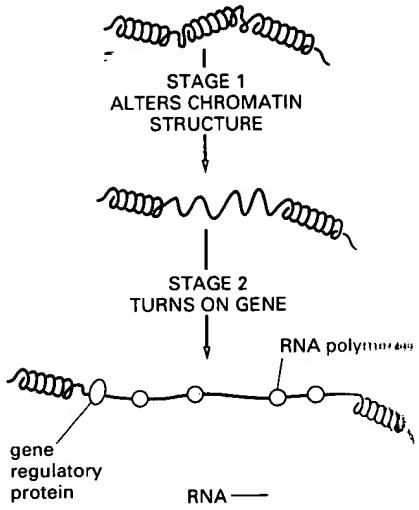


Figure 9–53 **The two stages postulated to be involved in some gene activations, such as that of the human globin gene cluster.** In stage 1 the structure of a large local region of chromatin is modified to decondense it in preparation for transcription. In stage 2 gene regulatory proteins (represented by a single protein in this simplified figure) bind to specific sites on the altered chromatin to induce RNA synthesis (transcription)

Table 9–2 The Expression of a Gene Transferred to a Mouse Generally Shows Chromosome Position Effects, with Different Levels of Gene Activity in Independently Derived Transgenic Animals

	Percent of Total mRNA in				Gene Copies per Cell
	Yolk Sac	Liver	Gut	Brain	
Endogenous gene	20	5	0.1	0	2
Transgenic animal 1	3.4	1	0.1	0	4
Transgenic animal 2	4.8	30	1.3	0	4
Transgenic animal 3	4.4	13	4.7	0	4
Transgenic animal 4	0.4	0.4	0	0	12

In these experiments, carried out with the mouse alpha-fetoprotein gene, the DNA fragment injected into the fertilized mouse egg included 14,000 nucleotide pairs of upstream (5′-flanking) sequence, where three enhancers that affect the expression of this gene are located. Hybridization was used to compare the level of mRNA produced by the injected gene to that normally produced by the endogenous mouse alpha-fetoprotein gene in the indicated fetal tissues. (Data from R.E. Hammer et al., *Science* 235:53–58, 1987.)

However, β-globin is expressed at high levels in erythroid cells regardless of the site of insertion, indicating that the LCR can override these position effects.

Although several proteins that specifically bind to the LCR have been identified, the mechanism that alters the chromatin structure of the entire β-globin locus is not known. Some ideas for how such changes may be brought about are discussed in the next section.

The Mechanisms That Form Active Chromatin Are Not Understood

The hypothetical model for globin activation outlined in Figure 9–53 implies that eucaryotes may contain sequence-specific DNA-binding proteins that function to decondense the chromatin in a local chromosomal domain that extends for tens of thousands of DNA nucleotide pairs. Alternatively, the observed differences in the chromatin structure of active genes could be an automatic consequence of the assembly of transcription factors or RNA polymerase (or both) onto a promoter rather than being a prerequisite for these events. We saw earlier that the assembly of the general transcription factors at promoters appears to be accompanied by changes in nucleosome distribution at the assembly site; perhaps this small perturbation can spread for long distances by some unknown propagation mechanism.

Whether or not eucaryotes turn out to have proteins that are specifically designed to decondense domains of chromatin, it is worth speculating on how proteins might accomplish this task. At present we can only guess at the mechanism. Three possibilities are outlined in Figure 9–54. These very different types of models indicate how far we are from understanding the transition from inactive to active chromatin.

Superhelical Tension in DNA Allows Action at a Distance [40]

One of the three models outlined in Figure 9–54 invokes topological changes in a closed loop of DNA double helix that can lead to the formation of DNA supercoils, a conformation that DNA adopts in response to *superhelical tension.* DNA supercoiling is most readily studied in small circular DNA molecules, such as the chromosomes of some viruses and plasmids. The same considerations apply, however, to any region of DNA bracketed by two ends that are unable to rotate freely—as, for example, in a loop of chromatin that is tightly clamped at its base.

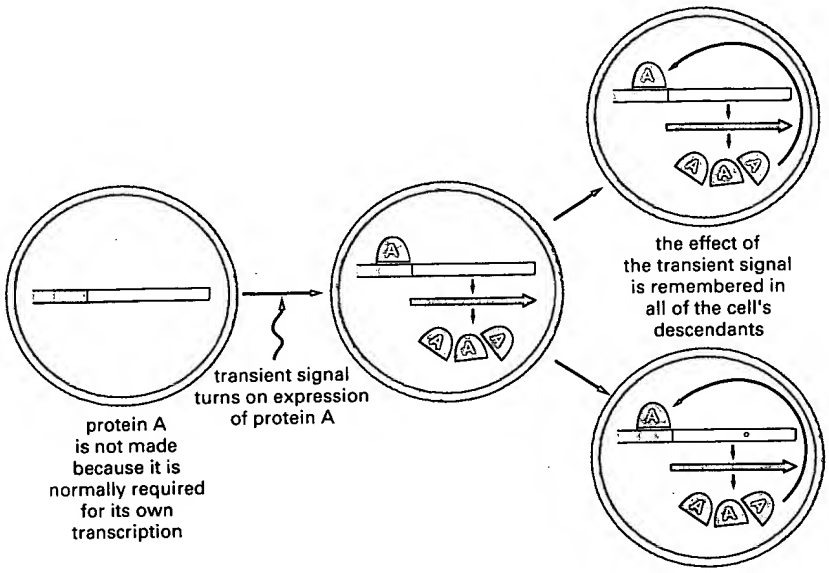


Figure 9–61 **Schematic diagram showing how a positive feedback loop can create cell memory.** Protein A is a gene regulatory protein that activates its own transcription. All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of the protein.

sor protein. In the absence of such interference, however, the lambda repressor both turns off production of the cro protein and turns on its own synthesis, and this *positive feedback loop* helps to maintain the prophage state. Positive feedback loops are a feature of many cell memory circuits (Figure 9–61).

Bacteriophage lambda illustrates an important general principle: a sophisticated pattern of inherited behavior can be achieved with only a few gene regulatory proteins that reciprocally affect one another's synthesis and activities. We know that variations of this simple strategy are used by eucaryotic cells to establish and maintain heritable patterns of gene transcription. Several gene regulatory proteins that are involved in establishing the *Drosophila* body plan (discussed in Chapter 21), for example, stimulate their own transcription, thereby creating a positive feedback loop that promotes their continued synthesis; at the same time these proteins repress the transcription of genes encoding other important gene regulatory proteins.

Expression of a Critical Gene Regulatory Protein Can Trigger Expression of a Whole Battery of Downstream Genes [45]

In general, a combination of multiple gene regulatory proteins, rather than a single protein, determines where and when a gene is transcribed in eucaryotes. But even if control is combinatorial, a single gene regulatory protein can be decisive in switching a cell from one developmental pathway or state of differentiation to another. A striking example comes from experiments on muscle cell differentiation *in vitro*.

A mammalian skeletal muscle cell is typically extremely large and contains many nuclei. It is formed by the fusion of many muscle precursor cells called *myoblasts*. The mature muscle cell is distinguished from other cells by a large number of characteristic proteins, including specific types of actin, myosin, tropomyosin, and troponin (all part of the contractile apparatus), creatine phosphokinase (for the specialized metabolism of muscle cells), and acetylcholine receptors (to make the membrane sensitive to nerve stimulation). In proliferating myoblasts these muscle-specific proteins and their mRNAs are absent or are present in very low concentrations. As myoblasts begin to fuse with one another, the corresponding genes are all switched on coordinately as part of a general transformation of the pattern of gene expression.

This entire program of muscle differentiation can be triggered in cultured skin fibroblasts and certain other cell types by introducing any one of a family

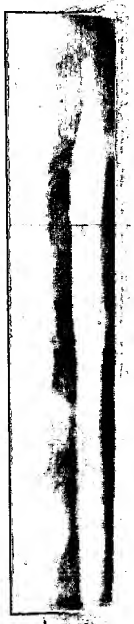


Figure 9–62 **The effect of expressing the MyoD protein in fibroblasts.** As shown in this immunofluorescence micrograph, skin fibroblasts from a chick embryo have been converted to muscle cells by the experimentally induced expression of the *myoD* gene. The fibroblasts were grown in culture and transfected three days earlier with a recombinant DNA plasmid containing the *myoD* coding sequence. Although only a few percent of the fibroblasts take up the DNA and produce the MyoD protein, these have fused to form elongated myotubes, which are stained here with an antibody that detects a muscle-specific protein. The muscle cells are intermixed with a confluent layer of fibroblasts, whose nuclei are barely visible in this micrograph. Control cultures transfected with another plasmid contain no muscle cells. (Courtesy of Stephen Tapscott and Harold Weintraub.)

helix-loop-helix proteins—the so-called *myogenic proteins* (MyoD, Myf5, or ogenin, for example)—normally expressed only in muscle cells (Figure 9–62). ding sites for these regulatory proteins can be detected in the regulatory DNA uences adjacent to many muscle-specific genes. From studies in transgenic e, it seems likely that MyoD and Myf5 act by turning on myogenin: if the ogenin gene is eliminated by targeted gene disruption, muscle cells fail to erentiate.

It is probable that the fibroblasts and other cell types that are converted to scle cells by myogenic proteins have already accumulated a number of gene ulatory proteins that can cooperate with the myogenic proteins to switch on scle-specific genes. In this view it is a specific *combination* of gene regulatory teins, rather than a single protein, that determines muscle differentiation. This a is consistent with the finding that some cell types fail to be converted to scle by myogenin or its relatives; these cells presumably have not accumulated other gene regulatory proteins required.

As we see next, combinatorial gene control has important implications for th the evolution and the development of multicellular organisms.

ombinatorial Gene Control Is the Norm in Eucaryotes [46]

have already discussed how multiple gene regulatory proteins can act in nbination to regulate the expression of an individual gene. But, as the example the myogenic proteins shows, combinatorial gene control means more than is; not only does each gene have many gene regulatory proteins to control it, each regulatory protein contributes to the control of many genes. Moreover,

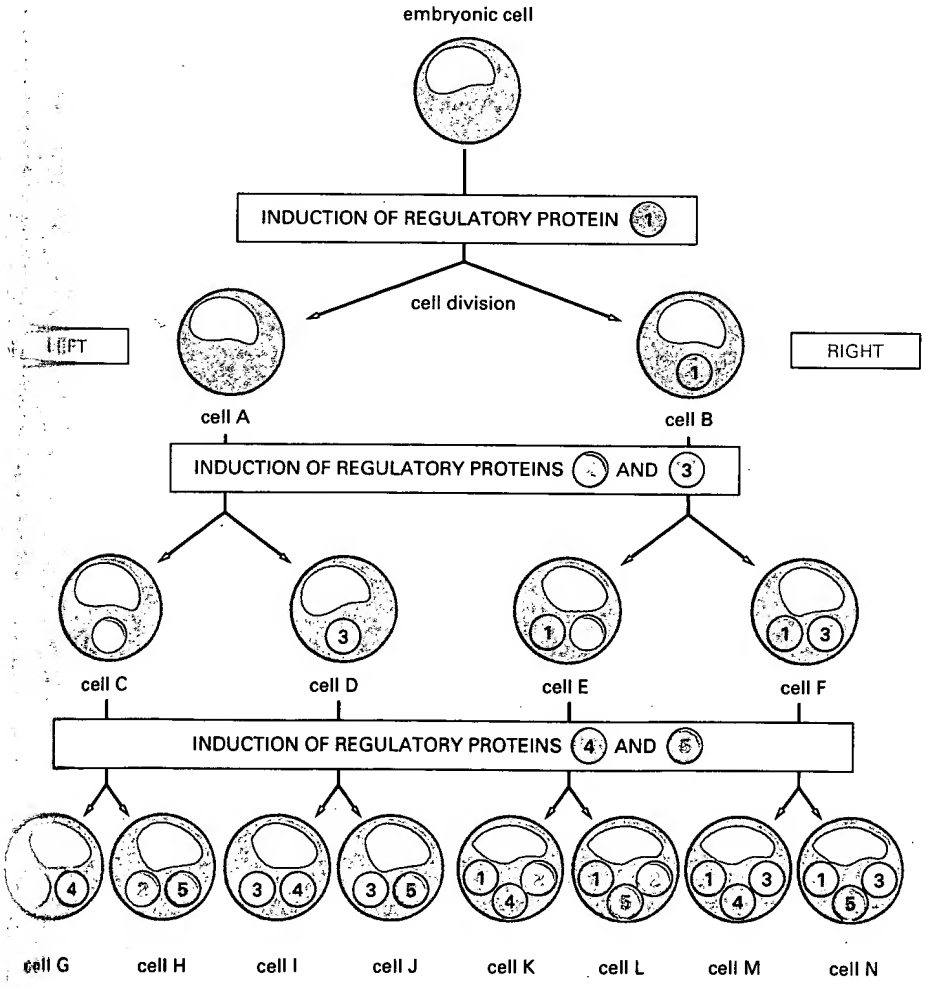


Figure 9–63 The importance of combinatorial gene control for development. A highly schematic scheme illustrating how combinations of a few gene regulatory proteins can generate many cell types during development. In this simple scheme a "decision" to make one of a pair of different gene regulatory proteins (shown as numbered circles) is made after each cell division. Sensing its relative position in the embryo, the daughter cell toward the left side of the embryo is always induced to synthesize the even-numbered protein of each pair, while the daughter cell toward the right side of the embryo is induced to synthesize the odd-numbered protein. The production of each gene regulatory protein is assumed to be self-perpetuating (thereby contributing to cell memory). Therefore, the cells in the enlarging clone contain an increasing number of regulatory proteins. Note that, in this purely hypothetical example, eight cell types (G through N) have been created with 5 different gene regulatory proteins. With continuation of such a scheme, more than 10,000 cell types could have been specified by only 25 different gene regulatory proteins.

although some gene regulatory proteins, like MyoD or myogenin, are specific to a single cell type, more typically production of a given gene regulatory protein is itself switched on in a variety of cell types, at several sites in the body, and at several times in development. This point is illustrated schematically in Figure 9–63, which shows how combinatorial gene control makes it possible to generate a great deal of biological complexity with relatively few gene regulatory proteins.

With combinatorial control, a given gene regulatory protein does not necessarily have a single, simply definable function as commander of a particular battery of genes or specifier of a particular cell type. Instead, it may serve many purposes that overlap with those of other gene regulatory proteins. These proteins can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event.

A consequence of combinatorial gene control is that the effect of adding a new gene regulatory protein to a cell will depend on the cell's past history, since this history will determine which gene regulatory proteins are already present. Thus during development a cell can accumulate a series of gene regulatory proteins that need not initially alter gene expression. When the final member of the requisite combination of gene regulatory proteins is added, however, the regulatory message is completed, leading to large changes in gene expression. Such a scheme, as we have seen, could explain how the addition of a single regulatory protein to a fibroblast can produce the dramatic transformation of the fibroblast into a muscle cell. It also can account for the important difference, discussed in Chapter 21, between the process of *cell determination,* where a cell becomes committed to a particular developmental fate, and the process of *cell differentiation,* where a committed cell expresses its specialized character. It is an essential feature of this scheme that once a gene regulatory protein has been made, it may act to maintain its own expression, thereby contributing to *cell memory,* which we discuss further below.

Combinatorial gene control also has an important consequence for evolution. Because gene regulatory proteins are not dedicated to a particular circuit or to a particular target gene, a subtle change in one gene regulatory protein can affect the expression pattern of many genes and thereby cause a substantial change in cell behaviors.

An Inactive X Chromosome Is Inherited [47]

We saw earlier how gene regulatory proteins can produce heritable patterns of gene expression in both procaryotic and eucaryotic cells. One possible mechanism, based on positive feedback in the control of gene expression, was illustrated in Figure 9–61. An additional mechanism operates only in eucaryotes, where long-range patterns of chromatin structure can be stably inherited. Perhaps the most dramatic example known is the inactivation of one of the two X chromosomes in female mammalian cells.

The X and Y chromosomes are the *sex chromosomes* of mammals: female cells contain two X chromosomes, while male cells contain one X and one Y chromosome. Presumably because a double dose of X-chromosome products would be lethal, the female cells have evolved a mechanism for permanently inactivating one of the two X chromosomes in each cell. In mice this occurs between the third and the sixth day of development, when, at random, one or other of the two X chromosomes in each cell becomes highly condensed into heterochromatin. This chromosome is seen in the light microscope during interphase as a distinct structure known as a *Barr body,* located near the nuclear membrane, and it replicates late in S phase. Most of its DNA is not transcribed. Because the inactive state of this X chromosome is faithfully inherited, every female is a mosaic composed of a mixture of clonal groups of cells in which only the paternally inherited X chro-

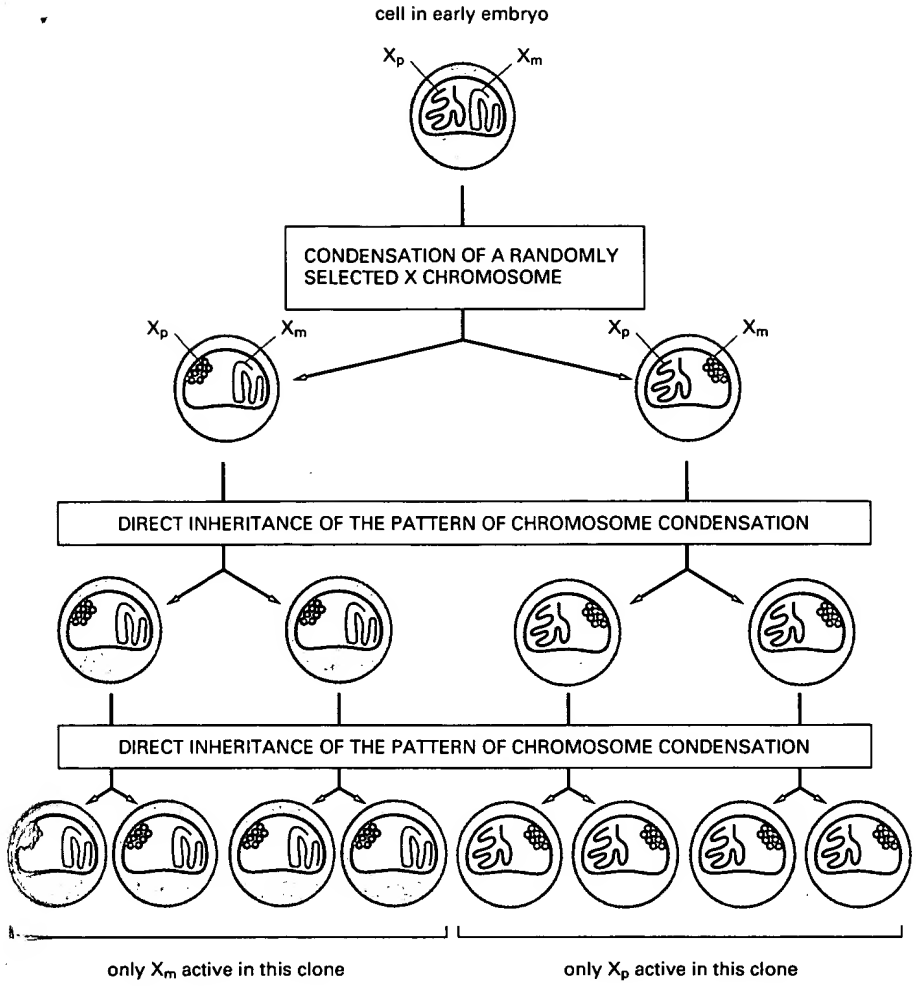


Figure 9–64 **X inactivation.** The clonal inheritance of a condensed inactive X chromosome that occurs in female mammals.

...osome (X_p) is active and a roughly equal number of clonal groups of cells in ...ich only the maternally inherited X chromosome (X_m) is active. In general, the ...ells expressing X_p and those expressing X_m are distributed in small clusters in ...he adult animal, reflecting the tendency of sister cells to remain close neighbors ...uring the later stages of embryonic development and growth (Figure 9–64).

The process that forms the condensed chromatin (the heterochromatin) in ...n X chromosome tends to spread continuously along the chromosome. This can ...e seen in studies with mutant animals in which one of the X chromosomes has ...ecome joined to a portion of an *autosome* (a nonsex chromosome). In such ...ybrid chromosomes regions of the autosome adjacent to an inactivated X chro...osome are often condensed into heterochromatin, causing the genes they con...in to be inactivated in a heritable way. This suggests that X-chromosome in...tivation occurs by a cooperative process that can be thought of as a chromatin "...rystallization" event that spreads linearly along the DNA from a nucleation site ...n the X chromosome. In fact, a unique *inactivation center* has been located ...enetically on the X chromosome: broken fragments of X chromosome do not ...ndergo inactivation unless they include this center.

Once the condensed chromatin structure is established on an X chromo...ome, some unknown process causes the structure to be faithfully inherited ...uring all subsequent replications of the DNA. The change is not absolutely per...anent, however, as the condensed X chromosome is reactivated in the forma...ion of germ cells in the female.

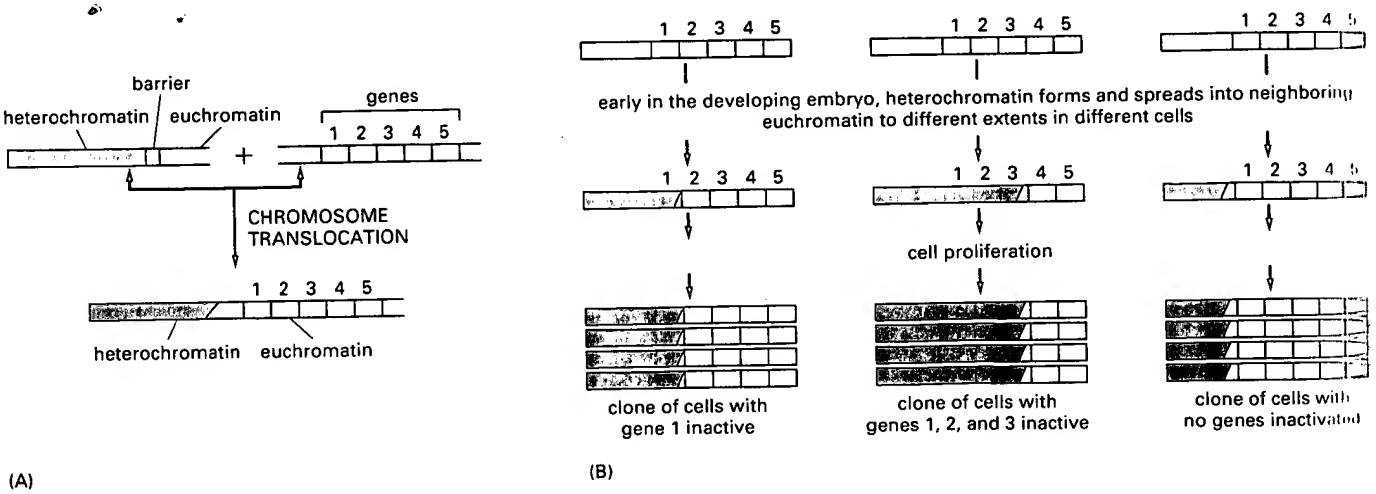


Figure 9–65 **Position-effect variegation in *Drosophila*.** (A) Heterochromatin (*red*) is normally prevented from spreading into adjacent regions of euchromatin (*green*) by special barrier sequences of unknown nature. In flies that inherit certain chromosomal translocations, however, this barrier is no longer present. (B) During the early development of such flies, the heterochromatin now spreads into neighboring chromosomal DNA, proceeding for different distances in different cells. The spreading soon stops, but the established pattern of heterochromatin is inherited, so that large clones of progeny cells are produced that have the same neighboring genes condensed into heterochromatin and thereby inactivated (hence the "variegated" appearance of some of these flies; see Figure 9–51B). This phenomenon shares many features with X-chromosome inactivation in mammals.

Drosophila and Yeast Genes Can Also Be Inactivated by Heritable Features of Chromatin Structure [48]

Earlier we discussed two examples of *position effects* on gene expression—one in *Drosophila* and one in yeast—that seem in many ways to be analogous to X-chromosome inactivation (see Figure 9–51). In both cases a specifically condensed form of chromatin prevents the expression of genes, and in both cases the condensed state of chromatin is heritable.

In flies with chromosomal rearrangements, breaking and rejoining events that place the middle of a region of heterochromatin next to a region of normal chromatin (*euchromatin*) tend to inactivate the nearby euchromatic genes. The situation is analogous to fusing a mammalian autosome to an inactive X chromosome, as just described, and the inactivation events are similarly patterned: the zone of inactivation spreads from the chromosome breakpoint to involve one or more adjacent genes. Moreover, while the extent of the spreading effect is different in different cells, the inactivated zone established in an embryonic cell is stably inherited by all of the cell's progeny (Figure 9–65). The example of position effect in yeast described previously in Figure 9–51 also shares some of the features of X-chromosome inactivation, including the spreading effect and the heritability of the condensed chromatin state.

It has not yet been proved that X-chromosome inactivation and the position effects in flies and yeast all occur by related mechanisms. Nevertheless, the parallels are striking. The recent identification and cloning of several *Drosophila* and yeast genes required for the position effects have provided an experimental entry point for exploring the molecular mechanisms involved. Figure 9–66 shows one hypothetical scheme that could, in principle, account for both the spreading effect and the heritable nature of the condensed chromatin state.

Regardless of its molecular basis, the packing of selected regions of the genome into condensed chromatin is a type of genetic regulatory mechanism that is not available to bacteria. The crucial feature of this uniquely eucaryotic form of gene regulation is the storing of the stable memory of gene states in an inherited chromatin structure rather than in a stable feedback loop of self-regulating gene regulatory proteins that can diffuse from place to place in the nucleus. Whether mechanisms of this type operate only to inactivate large regions of chromosomes or whether they can also operate at the level of one or a few genes is not known.

The Pattern of DNA Methylation Can Be Inherited When Vertebrate Cells Divide [49]

The nucleotides in DNA can be covalently modified, and in vertebrate cells the methylation of cytosine seems to provide an important mechanism for distin-

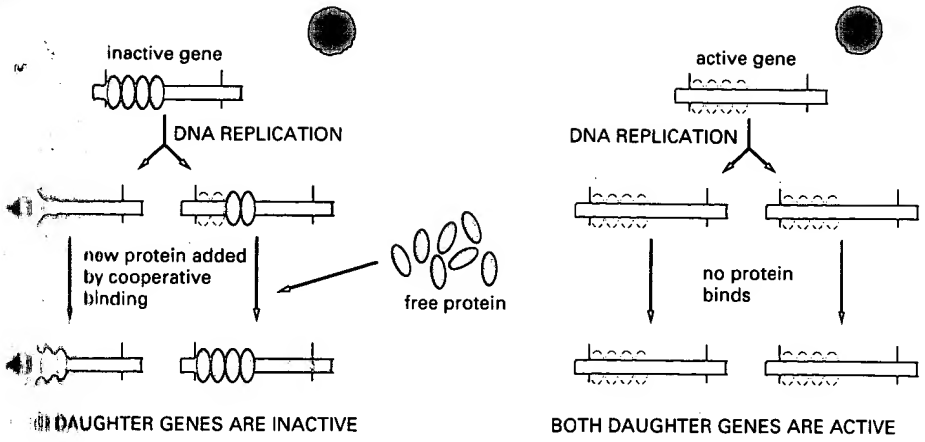


Figure 9–66 A general scheme that permits the direct inheritance f states of gene expression during DNA replication. In this hypothetical model, portions of a cooperatively bound cluster of chromosomal proteins are transferred directly from the parental DNA helix (*top left*) to both daughter helices. The inherited cluster then causes each of the daughter DNA helices to bind additional copies of the same proteins. Because the binding is cooperative, DNA synthesized from an identical parental DNA helix that lacks the bound proteins (*top right*) will remain free of them. If the bound proteins turn off gene transcription, then the inactive gene state will be directly inherited, as illustrated. If the cooperative protein binding requires specific DNA sequences, these events will be limited to specific gene control regions; if the binding can be propagated all along the chromosome, however, it could account for the spreading effect associated with the heritable chromatin states discussed in the text.

hing genes that are active from those that are not. The covalently modified *ethylcytosine (5-methyl C)* has the same relation to cytosine that thymine has ncil and likewise has no effect on base-pairing (Figure 9–67). The methylation vertebrate DNA is restricted to cytosine (C) nucleotides in the sequence CG, ch is base-paired to exactly the same sequence (in opposite orientation) on other strand of the DNA helix. Consequently, a simple mechanism permits existing pattern of **DNA methylation** to be inherited directly by the daughDNA strands. An enzyme called *maintenance methylase* acts preferentially on CG sequences that are base-paired with a CG sequence that is *already* thylated. As a result, the pattern of DNA methylation on the parental DNA and will act as a template for the methylation of the daughter DNA strand, ing this pattern to be inherited directly following DNA replication (Figure 9–

Bacteria produce enzymes that are useful for studying methylation in vertete cells. They use the methylation of either an A or a C at a specific site to otect themselves from the action of their own restriction nucleases. The restricn nuclease HpaII, for example, cuts the sequence CCGG but fails to cleave it the central C is methylated. Thus the susceptibility of a DNA molecule to cleavby HpaII can be used to detect whether CG sequences at specific DNA sites methylated. The inheritance of methylation patterns can be studied in verbrate cells in culture by first using bacterial methylating enzymes to introduce thyl groups on cytosines and then using bacterial restriction nucleases to folw the inheritance of these groups. The enzyme used to introduce 5-methyl C es into specific CG sequences is the HpaII-methylase that normally protects bacterium against its own HpaII restriction nuclease. If this enzyme is used methylate the central C in the sequence CCGG on a cloned DNA molecule that introduced into cultured vertebrate cells, the maintenance methylase can be own to work as expected: each individual methylated CG is generally retained rough many cell divisions, whereas unmethylated CG sequences remain nmethylated.

The maintenance methylase explains the automatic inheritance of 5-methyl nucleotides, but since it normally does not methylate fully unmethylated DNA, leaves unanswered the question of how the methyl group is first added in a rtebrate organism. If a fully unmethylated DNA molecule is injected into a rilized mouse egg, methyl groups will be added to nearly every CG site (an portant exception will be described below). This is presumed to reflect the esence of a novel *establishment methylase* activity in the egg. As we shall see, *novo* methylation can also occur during the differentiation of specialized cell pes, although it is not known how it occurs.

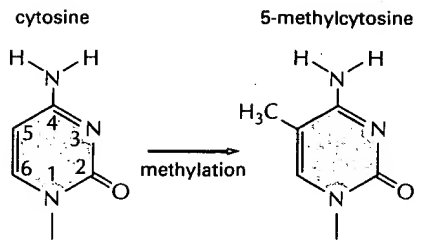
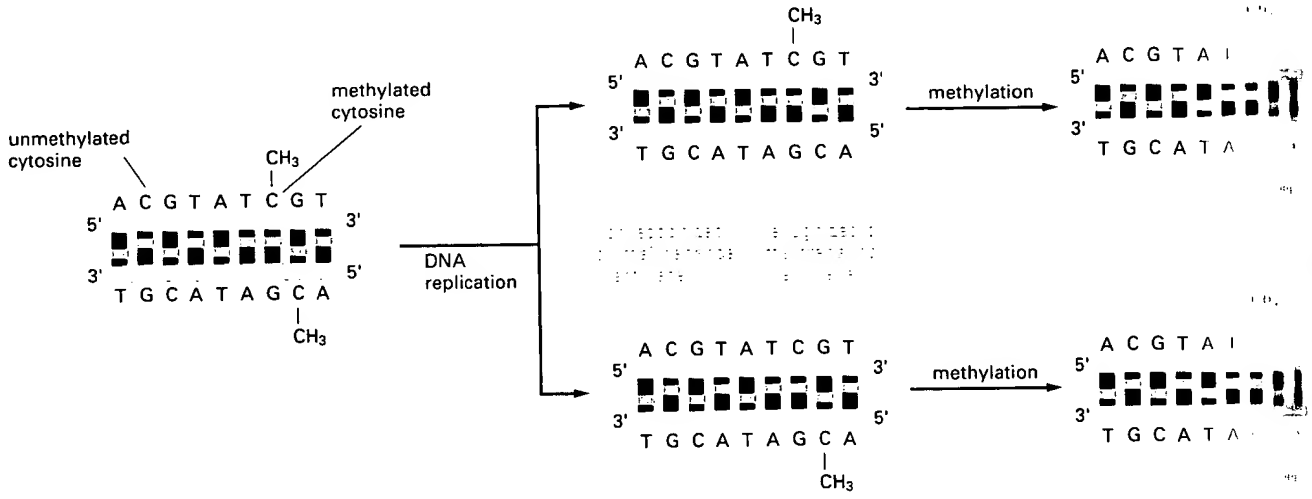


Figure 9–67 Formation of 5-methyl-cytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates this event is confined to selected cytosine (C) nucleotides located in the sequence CG.



How DNA methylation patterns are faithfully inherited. In vertebrate DNAs a large fraction of the cytosine nucleotides in the sequence CG are methylated (see Figure 9–67). Because of the action of a methyl-directed methylase enzyme (the maintenance methylase), once a pattern of DNA methylation is established, each site of methylation is inherited in the progeny DNA, as shown. This means that changes in DNA methylation patterns will be perpetuated in all of the progeny of a cell.

DNA Methylation Reinforces Developmental Decisions in Vertebrate Cells [50]

Although DNA methylation was once proposed to play a dominant part in generating different mammalian cell types, it is now viewed as having a more subtle role. In some invertebrates, including *Drosophila*, DNA methylation does not occur, yet the control of gene expression and the diversification of cell types appear to be similar in *Drosophila* and vertebrates. Several observations are consistent with the idea that DNA methylation in vertebrates is associated with gene inactivation but that it usually only reinforces decisions that are first brought about by other mechanisms. Thus tests with the HpaII restriction nuclease indicate that in general the DNA of inactive genes is more heavily methylated than that of active genes. When an inactive gene that contains methylated DNA is turned on during the course of normal development, however, it generally loses most of its methyl groups only after the gene has been activated. Conversely, the female X chromosome, discussed above, is first condensed and inactivated and only later acquires an increased level of methylation on some of its genes.

What, then, does methylation do, and why is it useful to the organism? There are at least two important clues. First, the DNA corresponding to a muscle-specific actin gene can be prepared in both its fully methylated and its fully unmethylated form. When these two versions of the gene are introduced into cultured muscle cells, both are transcribed at the same high rate. When, however, they are introduced into fibroblasts, which normally do not transcribe the gene, the unmethylated gene is transcribed at a low rate, whereas neither the exogenously added methylated gene nor the endogenous gene of the fibroblast (which is also methylated) is transcribed at all. Second, biochemical experiments have identified a vertebrate protein that binds tightly to DNA that contains clustered 5-methyl C nucleotides. The binding of this protein is thought to package the methylated DNA in a way that makes it unusually resistant to the transcriptional activation machinery. These two observations suggest that DNA methylation is used in vertebrates mainly to ensure that once a gene is turned off, it stays completely off (Figure 9–69).

Experiments designed to test whether a DNA sequence that is transcribed at high levels in one vertebrate cell type is transcribed at all in another have demonstrated that rates of gene transcription can differ between two cell types by a factor of more than 10^6. Thus unexpressed vertebrate genes are much less "leaky" in terms of transcription than are unexpressed genes in bacteria, in which the

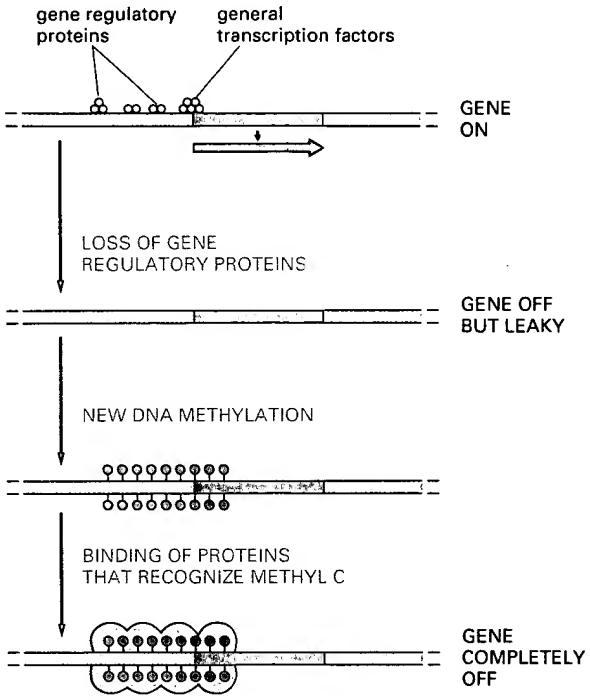


Figure 9–69 **How DNA methylation may help turn off genes.** The binding of gene regulatory proteins and general transcription factors near an active promoter prevents DNA methylation by some unknown mechanism. If most of these sequence-specific DNA-binding proteins dissociate, however, as generally occurs when a gene is turned off, the DNA becomes methylated, which enables other proteins to bind, and these shut down the gene completely.

known differences in transcription rates between expressed and unexpressed gene states are about 1000-fold. DNA methylation of unexpressed vertebrate genes may account for at least part of this difference. In addition, as we shall see next, DNA methylation is required for at least one special type of cellular memory.

Genomic Imprinting Requires DNA Methylation [51]

Mammalian cells are diploid, containing one set of genes inherited from the father and one set from the mother. In a few cases the expression of a gene has been found to depend on whether it is inherited from the mother or the father. This phenomenon is called **genomic imprinting.** Although not originally discovered this way, genomic imprinting has been dramatically illustrated in experiments with transgenic mice. It is possible, for example, to make transgenic mice in which one of the two normal copies of the gene coding for *insulinlike growth factor-2 (IGF-2)* has been inactivated by mutation. These heterozygous mice develop normally if it is the maternally derived *Igf-2* gene that is defective, but if the paternally derived *Igf-2* gene is defective, they are stunted, growing to less than half the size of normal mice. Further analysis of these and normal mice provided an explanation. In both the transgenic and the wild-type mice only the paternally derived *Igf-2* gene is transcribed, while the maternally derived gene is silent. The maternally derived gene in this case is said to be *imprinted.*

Although the mechanism of imprinting is uncertain, it seems very likely that DNA methylation is involved. Thus, in transgenic mice defective in the maintenance methylase, the imprinting of the maternal *Igf-2* gene does not occur, implying that the mechanism that distinguishes between the paternal and maternal copies of the *Igf-2* gene requires DNA methylation. Interestingly, the mice defective in the maintenance methylase die as young embryos. This could result from defective imprinting, but it is also conceivable that a failure to reinforce developmental decisions by methylation is the primary defect, leading to leaky transcription of the many thousands of genes that are normally turned off in each differentiated cell.

CG-rich Islands Are Associated with About 40,000 Genes in Mammals [52]

Because of the way DNA repair enzymes work, methylated C nucleotides in the genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U, which is not normally present in DNA and thus is recognized easily by the DNA repair enzyme uracil DNA glycosylase, excised, and then replaced with a C (discussed in Chapter 6). But accidental deamination of a 5-methyl C cannot be repaired in this way, for the deamination product is a T and so indistinguishable from the other, nonmutant T nucleotides in the DNA. Although a special repair system exists to remove these mutant Ts (see p. 250), many of the deaminations escape detection, so that those C nucleotides in the genome that are methylated tend to mutate to T over evolutionary time.

During the course of evolution, more than three out of every four CGs have been lost in this way, leaving vertebrates with a remarkable deficiency of this dinucleotide. The CG sequences that remain are very unevenly distributed in the genome; they are present at 10 to 20 times their average density in selected regions, called **CG islands,** that are 1000 to 2000 nucleotide pairs long. These islands, with some important exceptions, seem to remain unmethylated in all cell types. They are thought to surround the promoters of the so-called *housekeeping genes*—those genes that code for the many proteins that are essential for cell viability and are therefore expressed in most cells (Figure 9–70). In addition, many *tissue-specific genes,* which code for proteins needed only in selected types of cells, are also associated with CG islands.

The distribution of CG islands can be explained if we assume that CG methylation was adopted in vertebrates as a way of hindering the initiation of transcription in inactive segments of the genome (Figure 9–71). In the germ line of vertebrates—the cell lineage giving rise to eggs and sperm—most of the genome is inactive and methylated. Over long periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through accidental deamination events that were not correctly repaired. The CG sequences in the regions surrounding the promoters of many genes, however, including all housekeeping genes, are kept demethylated in cells of the germ line, and so they can be readily repaired after spontaneous deamination events. Such regions are preserved as CG islands.

The mammalian genome (about 3×10^9 nucleotide pairs) contains an estimated 40,000 CG islands. Most of the islands mark the 5′ ends of a transcription unit and thus, presumably, a gene. It is possible to clone specifically the DNA surrounding the CG islands, and this technique provides a convenient way of finding new genes.

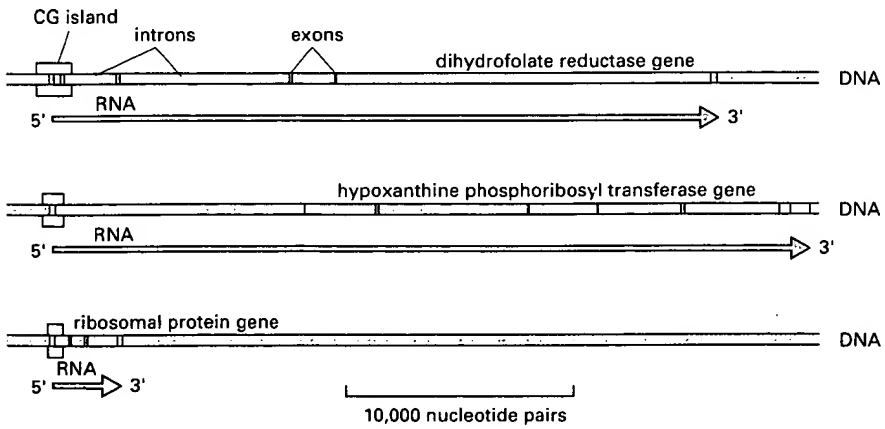


Figure 9–70 **The CG islands surrounding the promoter in three mammalian housekeeping genes.** The *yellow* boxes show the extent of each island. Note also that, as for most genes in mammals, the exons (*dark red*) are very short relative to the introns (*light red*). (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987.)

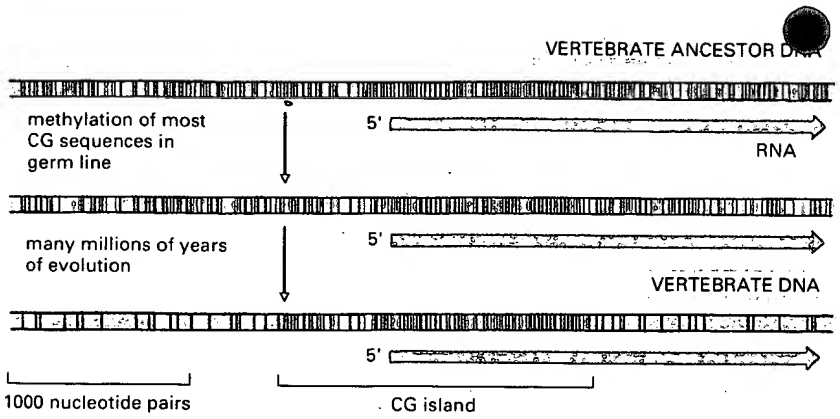


Figure 9–71 **A mechanism to explain both the marked deficiency of CG sequences and the presence f CG islands in vertebrate genomes.** A *black line* marks the location of an unmethylated CG dinucleotide in the DNA sequence, while a *red line* marks the location of a methylated CG dinucleotide.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides, endowing the cell with a memory of its developmental history. Procaryotes and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms, some of which may be relevant to the creation of specialized cell types in higher eucaryotes. One such mechanism involves a competitive interaction between two (or more) gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory.

In eucaryotes gene transcription is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be expressed in a variety of circumstances and typically is involved in the regulation of many genes.

In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also utilized by eucaryotic cells to regulate gene expression. In vertebrates DNA methylation also plays a part, mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms.

Posttranscriptional Controls

Although controls on the initiation of gene transcription are the predominant form of regulation for most genes, other controls can act later in the pathway from RNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls,** which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control,* for many genes they are crucial. It seems that every step in gene expression that could be controlled in principle is likely to be regulated under some circumstances for some genes.

We consider the varieties of posttranscriptional regulation in temporal order, according to the sequence of events that might be experienced by an RNA molecule after its transcription has begun (Figure 9–72).

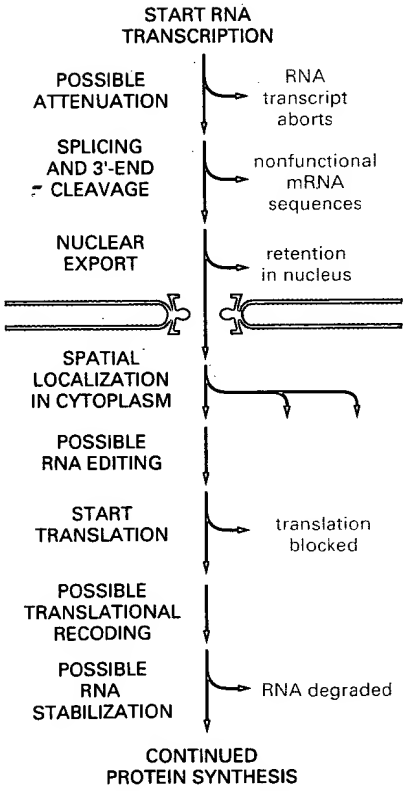


Figure 9–72 **Possible posttranscriptional controls on gene expression.** Only a few of these controls are likely to be used for any one gene.

Transcription Attenuation Causes the Premature Termination of Some RNA Molecules [53]

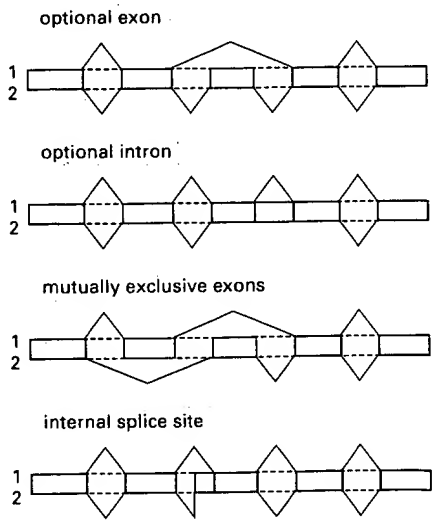
In bacteria the expression of certain genes is inhibited by premature termination of transcription, a phenomenon called **transcription attenuation.** In some of these cases the nascent RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription. When the gene product is required, regulatory proteins bind to the nascent RNA chain and interfere with attenuation, allowing the transcription of a complete RNA molecule.

In eucaryotes transcription attenuation can occur by a number of distinct mechanisms. In both adenovirus and HIV (the human AIDS virus), for example, the proteins that assemble at the promoter seem to determine whether or not the polymerase will be able to pass through specific sites of attenuation downstream. These proteins can differ from one cell type to the next, and the cell can control the degree of attenuation for particular genes.

Alternative RNA Splicing Can Produce Different Forms of a Protein from the Same Gene [54]

As discussed in Chapter 8, many genes are first transcribed as long mRNA precursors that are then shortened by a series of processing steps to produce the mature mRNA molecule. One of these steps is *RNA splicing,* in which the intron sequences are removed from the mRNA precursor. Often a cell can splice the primary transcript in different ways and thereby make different polypeptide chains from the same gene—a process called **alternative RNA splicing** (Figure 9–73). A substantial proportion of higher eucaryotic genes produce multiple proteins in this way. When different splicing possibilities exist at several positions in the transcript, a single gene can produce dozens of different proteins. Usually, however, the splice alternatives are more limited, and only a few kinds of proteins are synthesized from each transcription unit.

In some cases alternative RNA splicing occurs because there is an "intron sequence ambiguity": the standard spliceosome mechanism for removing intron sequences (discussed in Chapter 8) is unable to distinguish cleanly between two or more alternative pairings of 5′ and 3′ splice sites, so that different choices are made haphazardly on different occasions. Where such *constitutive alternative splicing* occurs, several versions of the protein encoded by the gene are made in all cells in which the gene is expressed.



Four patterns of alternative RNA splicing. In each a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark the sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences that are included in only one of the mRNAs; these boxes are joined by *red lines* to indicate where intron sequences (*yellow*) are removed. (Adapted with permission from A. Andreadis, J. Gallego, and B. Nadal-Ginard, *Annu. Rev. Cell Biol.* 3:207–242, 1987. © Annual Reviews, Inc.)

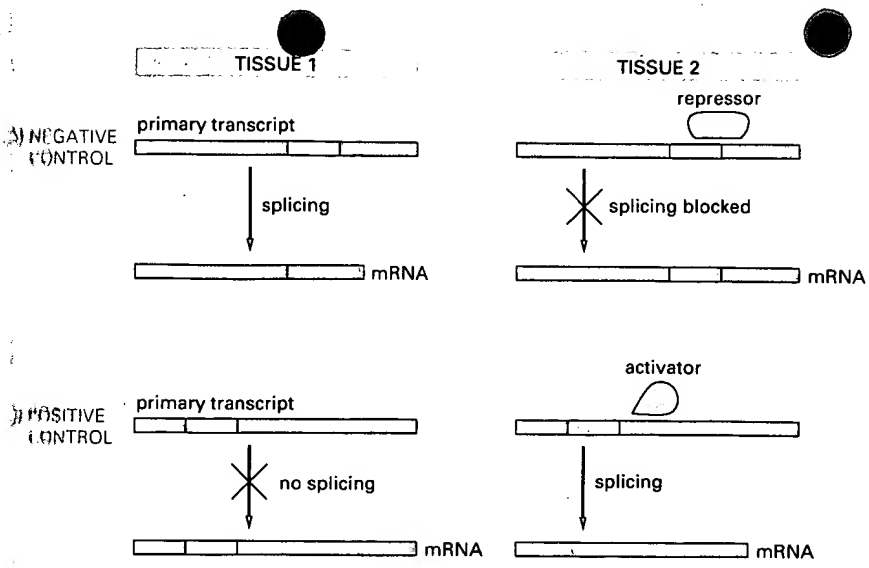


Figure 9–74 Negative and positive control of alternative RNA splicing. (A) Negative control, in which a repressor protein binds to the primary RNA transcript in tissue 2, thereby preventing the splicing machinery from removing an intron sequence. (B) Positive control, in which the splicing machinery is unable to remove a particular intron sequence without assistance from an activator protein.

In many cases, however, alternative RNA splicing is *regulated* rather than constitutive. In the simplest examples regulated splicing is used to switch from production of a nonfunctional protein to the production of a functional one. Transposase that catalyzes the transposition of the *Drosophila* P element, for example, is produced in a functional form in germ cells and a nonfunctional form in somatic cells of the fly, allowing the P element to spread throughout the genome of the fly without causing damage in somatic cells. The difference in transposon activity has been traced to the presence of an intron sequence in the transposase RNA that is removed only in germ cells.

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that directs the splicing machinery to an otherwise overlooked splice site (Figure 9–74). In the case of the *Drosophila* transposase, the key splicing event is blocked in somatic cells by negative regulation.

In addition to switching from the production of a functional protein to the production of a nonfunctional one, the regulation of RNA splicing can generate different versions of a protein in different cell types, according to the needs of the cell. The tyrosine protein kinase encoded by the *src* proto-oncogene, for example, is produced in a specialized form in nerve cells by this mechanism (Figure 9–75). Cell-type-specific forms of many other proteins are produced in the same

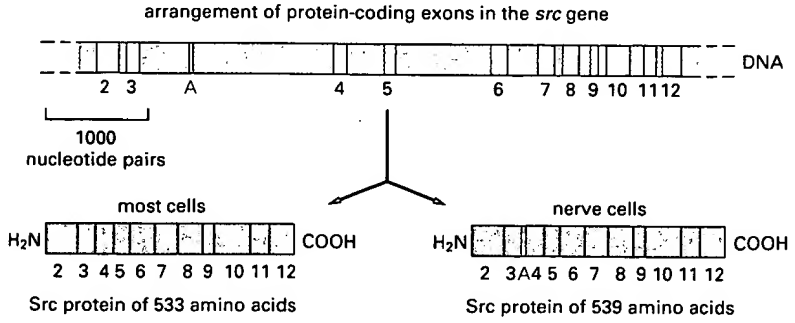


Figure 9–75 Regulated alternative RNA splicing produces cell-type-specific forms of a gene product. Here two slightly different tyrosine protein kinases are produced from the *src* gene because exon sequence A is included only in nerve cells. The neural form of the Src protein contains an extra site for phosphorylation and is also thought to have a higher specific activity. Only the protein-coding exons (*colored*) are shown in this diagram (exon 1, which forms the 5′ leader on the mRNA, is not shown). (After J.B. Levy et al., *Mol. Cell Biol.* 7:4142–4145, 1987.)

Sex Determination in *Drosophila* Depends on a Regulated Series of RNA Splicing Events [55]

In *Drosophila* the primary signal for determining whether the fly develops as a male or female is the X chromosome/autosome ratio. Individuals with an X chromosome/autosome ratio of 1 (normally two X chromosomes and two sets of autosomes) develop as females, while those with a ratio of 0.5 (normally one X chromosome and two sets of autosomes) develop as males. This ratio is somehow assessed early in development and is remembered by each cell thereafter. Three crucial gene products are involved in transmitting information about this ratio to the many other genes that specify male and female characteristics (Figure 9–76). As explained in Figure 9–77, sex determination in *Drosophila* depends on a cascade of regulated RNA splicing events that involves these three gene products.

Drosophila sex determination provides the best-understood example of a regulatory cascade based on RNA splicing. It is not clear why the fly should use this strategy. Other organisms (the nematode, for example) use an entirely different scheme for sex determination—one based on transcriptional and translational controls. Moreover, the *Drosophila* male-determination pathway requires that a number of nonfunctional RNA molecules be continually produced, which seems unnecessarily wasteful. One speculation is that this RNA-splicing cascade is an ancient control device, left over from a stage of evolution where RNA was the predominant biological molecule and controls of gene expression had to be based almost entirely on RNA-RNA interactions.

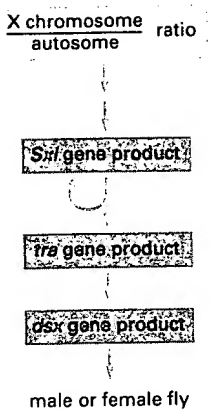


Figure 9–76 **Sex determination in *Drosophila*.** The gene products shown act in a sequential cascade to determine the sex of the fly according to the X chromosome/autosome ratio. The genes are called *sex-lethal (Sxl)*, *transformer (tra)*, and *doublesex (dsx)* because of the phenotypes that result when the gene is inactivated by mutation. The function of these gene products is to transmit the information about the X chromosome/autosome ratio to the many other genes that are involved in creating the sex-related phenotypes. These other genes function as two alternative sets: those that specify female features and those that specify male features (see Figure 9–77).

A Change in the Site of RNA Transcript Cleavage and Poly-A Addition Can Change the Carboxyl Terminus of a Protein [56]

In eucaryotes the 3′ end of an mRNA molecule is not determined by the termination of RNA synthesis by the RNA polymerase as it is in bacteria. Instead, it is determined by an RNA cleavage reaction that is catalyzed by additional factors while the transcript is elongating (see Figure 8–49). A cell can control the site of this cleavage so as to change the carboxyl terminus of the resultant protein (which is encoded by the 3′ end of the mRNA).

A well-studied example is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes. Early in the life history of a B lymphocyte, the antibody it produces is anchored in the plasma membrane, where it serves as a receptor for antigen. Antigen stimulation causes these cells to multiply and to start secreting their antibody. The secreted form of the antibody is identical to the membrane-bound form except at the extreme carboxyl terminus. In this part of the protein the membrane-bound form has a long string of hydrophobic amino acids that traverses the lipid bilayer of the membrane, whereas the secreted form has a much shorter string of hydrophilic amino acids. The switch from membrane-bound to secreted antibody therefore requires a different nucleotide sequence at the 3′ end of the mRNA; this difference is generated through a change in the length of the primary RNA transcript caused by a change in the site of RNA cleavage, as described in Figure 9–78.

The Definition of a Gene Has Had to Be Modified Since the Discovery of Alternative RNA Splicing [57]

The discovery that eucaryotic genes usually contain introns and that their coding sequences can be put together in more than one way raised new questions about the definition of a gene. A gene was first clearly defined in molecular terms in the early 1940s from work on the biochemical genetics of the fungus *Neurospora*. Until then, a **gen** had been defined operationally as a region of the ge-

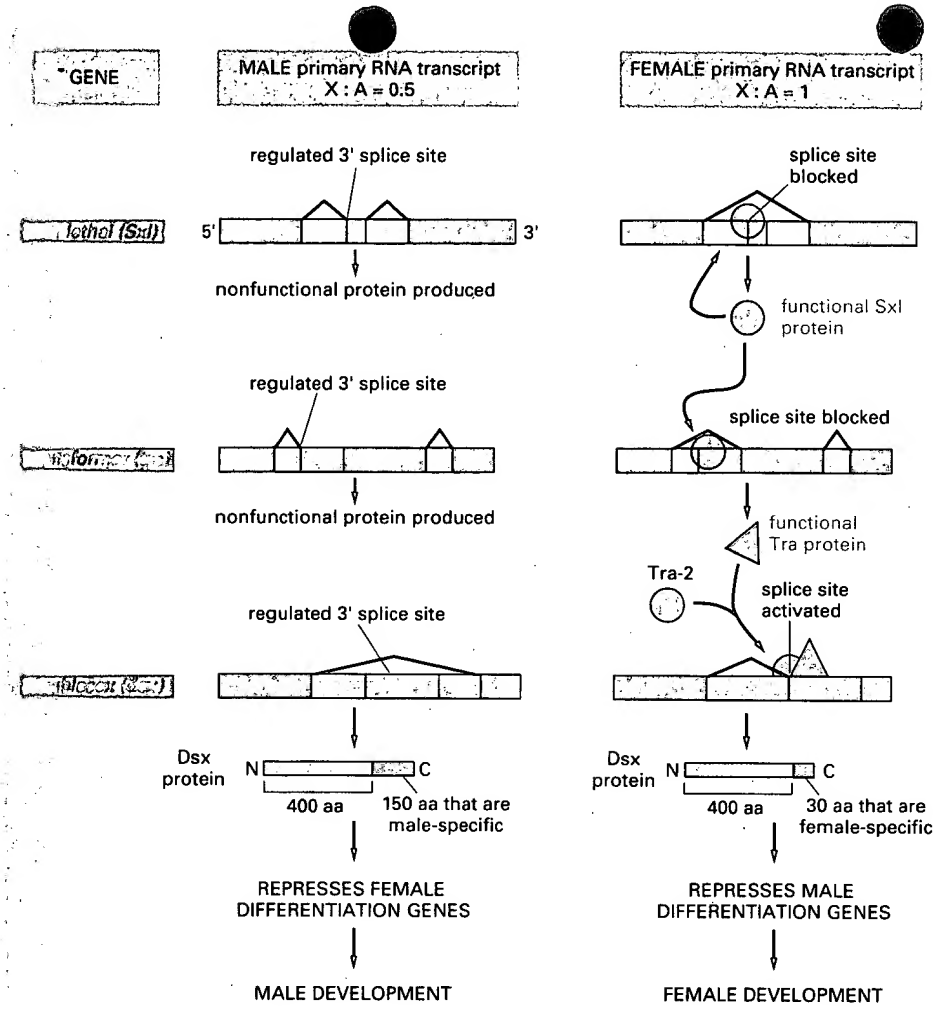
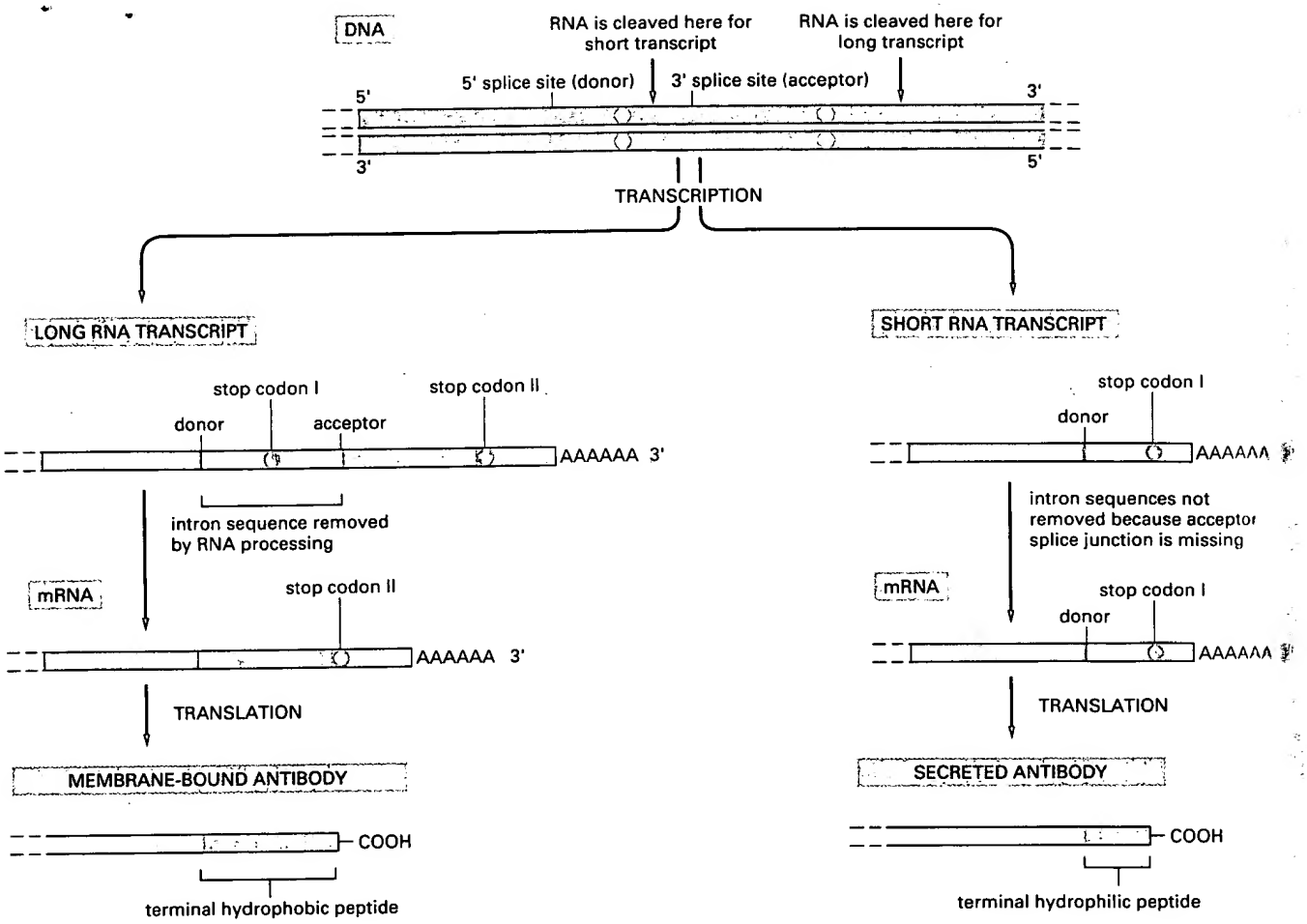


Figure 9–77 **The cascade of changes in gene expression that determines the sex of a fly depends on alternative RNA splicing.** An X chromosome/autosome ratio of 0.5 results in male development. Male is the "default" pathway in which the *Sxl* and *tra* genes are both transcribed, but the RNAs are spliced constitutively to produce only nonfunctional RNA molecules, and the *dsx* transcript is spliced to produce a protein that turns off the genes that specify female characteristics. An X chromosome/autosome ratio of 1 triggers the female differentiation pathway in the embryo by transiently activating a promoter within the *Sxl* gene that causes a functional Sxl protein to be synthesized. Sxl is a splicing regulatory protein with two sites of action: (1) it binds to a constitutively produced *Sxl* RNA transcript, causing a female-specific splice that continues the production of a functional Sxl protein, and (2) it binds to the constitutively produced *tra* RNA and causes an alternative splice of this transcript, which now produces an active Tra regulatory protein. The Tra protein acts with the constitutively produced Tra-2 protein to produce the female-specific spliced form of the *dsx* transcript; this encodes the female form of the Dsx protein, which turns off the genes that specify male features. The components in this pathway were all initially identified through the study of *Drosophila* mutants that are altered in their sexual development. The *dsx* gene, for example, derives its name (*doublesex*) from the observation that a fly lacking this gene product expresses both male- and female-specific features. Note that, whereas both the Sxl and Tra proteins bind to specific RNA sites, Sxl is a repressor that acts negatively to block a splice, whereas the Tra proteins are activators that act positively to induce a splice (see Figure 9–74).

nome that segregates as a single unit during meiosis and gives rise to a definable phenotypic trait, such as a red or a white eye in *Drosophila* or a round or wrinkled seed in peas. The work on *Neurospora* showed that most genes correspond to a region of the genome that directs the synthesis of a single enzyme. This led to the hypothesis that one gene encodes one polypeptide chain. The hypothesis proved fruitful for subsequent research; and, as more was learned about the mechanism of gene expression in the 1960s, a gene became identified as that stretch of DNA that was transcribed into the RNA coding for a single polypeptide chain (or a single structural RNA such as a tRNA or an rRNA molecule). The discovery of split genes in the late 1970s could be readily accommodated by the original definition of a gene, provided that a single polypeptide chain was specified by the RNA transcribed from any one DNA sequence. But it is now clear that many DNA sequences in higher eucaryotic cells produce two or more distinct proteins by means of alternative RNA splicing. How then is a gene to be defined?

In those relatively rare cases in which two very different eucaryotic proteins are produced from a single transcription unit, the two proteins are considered to be produced by distinct genes that overlap on the chromosome. It seems unnecessarily complex, however, to consider most of the protein variants produced by alternative RNA splicing as being derived from overlapping genes. A more sensible alternative is to modify the original definition to include as a gene any DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (*protein isoforms*). This definition of a gene also accommodates those DNA sequences that encode protein variants produced by posttranscriptional processes other than RNA splicing, such as ribosomal frameshifting and RNA editing, which we discuss later.



RNA Transport from the Nucleus Can Be Regulated [58]

An average primary RNA transcript seems to be at most 10 times longer than the mature mRNA molecule generated from it by RNA splicing. Yet it has been estimated that only about one-twentieth of the total mass of RNA made ever leaves the nucleus. It seems, therefore, that a substantial fraction of the primary transcripts (perhaps half) may be completely degraded in the nucleus without ever generating an RNA molecule that reaches the cytoplasm. The discarded RNAs may consist of sequences that are never made into an mRNA molecule; on the other hand, some may represent potential mRNA molecules that are functional in some cell types but in others fail to get delivered to the cytoplasm. This might be either because they are selectively targeted for intranuclear degradation or because their exit from the nucleus is selectively blocked.

Although there is very little solid evidence for either form of control, each of them remains a real possibility. In particular, RNA export through the nuclear pores is an active process, and for most RNAs it requires a specific nucleotide cap at the 5′ end of the RNA molecule and a poly-A tail at the 3′ end (discussed in Chapter 8). A requirement of this type makes sense, since it keeps junk RNA fragments (such as the intron sequences removed by RNA splicing) out of the cytosol. But having the proper types of ends is not enough for transport: each mRNA precursor molecule remains tethered to sites inside the nucleus until all of the spliceosome components have dissociated from it (see Figure 8–54). Therefore, any mechanism that prevents the completion of RNA splicing on particular RNA molecules could, in principle, block the exit of those RNAs from the nucleus.

Figure 9–78 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound. In unstimulated B lymphocytes (*left*) a long RNA transcript is produced, and the intron sequence near its 3′ end is removed by RNA splicing to give rise to an mRNA molecule that codes for a membrane-bound antibody molecule. In contrast, after antigen stimulation (*right*) the primary RNA transcript is cleaved upstream from the splice site in front of the last exon sequence. As a result, some of the intron sequence that is removed from the long transcript remains as coding sequence in the short transcript. These are the sequences that encode the hydrophilic carboxyl-terminal portion of the secreted antibody molecule.

Some mRNAs Are Localized to Specific Regions of the Cytoplasm [59]

Once a newly made eucaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it encounters ribosomes that translate it into a polypeptide chain. If the mRNA encodes a protein that is destined to be secreted or expressed on the cell surface, it will be directed to the endoplasmic reticulum (ER) by a signal sequence at the protein's amino terminus; the signal sequence will be recognized as soon as it emerges from the ribosome by components of the cell's protein-sorting apparatus. This apparatus then directs the entire complex of ribosome, mRNA, and nascent protein to the membrane of the ER, where the remainder of the polypeptide chain is synthesized, as discussed in Chapter 12. In other cases the entire protein is synthesized by free ribosomes in the cytosol, and signals in the completed polypeptide chain may then direct the protein to other sites in the cell.

In still other cases, however, mRNAs are directed to specific intracellular locations by signals in the mRNA sequence itself, before the sequence has been translated into an amino acid sequence. The signal is typically located in the 3′ *untranslated region (UTR)* of the mRNA molecule—a region that extends from the stop codon, which terminates protein synthesis, to the start of the poly-A tail (see Figure 9–78). A striking example is seen in the *Drosophila* egg, where the mRNA encoding the bicoid gene regulatory protein is attached to the cortical cytoskeleton at the anterior tip of the developing egg. When the translation of this mRNA is triggered by fertilization, a gradient of the bicoid protein is generated that plays a crucial part in directing the development of the anterior part of the embryo (shown in Figure 9–39 and discussed in more detail in Chapter 21). Some mRNAs in somatic cells are localized in a similar way. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3′ UTR signal, presumably because it is advantageous for the cell to position its mRNAs close to the sites where the protein produced from the mRNA is required. This form of posttranscriptional gene regulation, where mRNA is specifically localized to one part of the cell, has been recognized only recently, and it is still unclear how many mRNAs are localized in this way. The role of the UTR in localizing mRNAs to a particular region of the cytoplasm is illustrated in Figure 9–79.

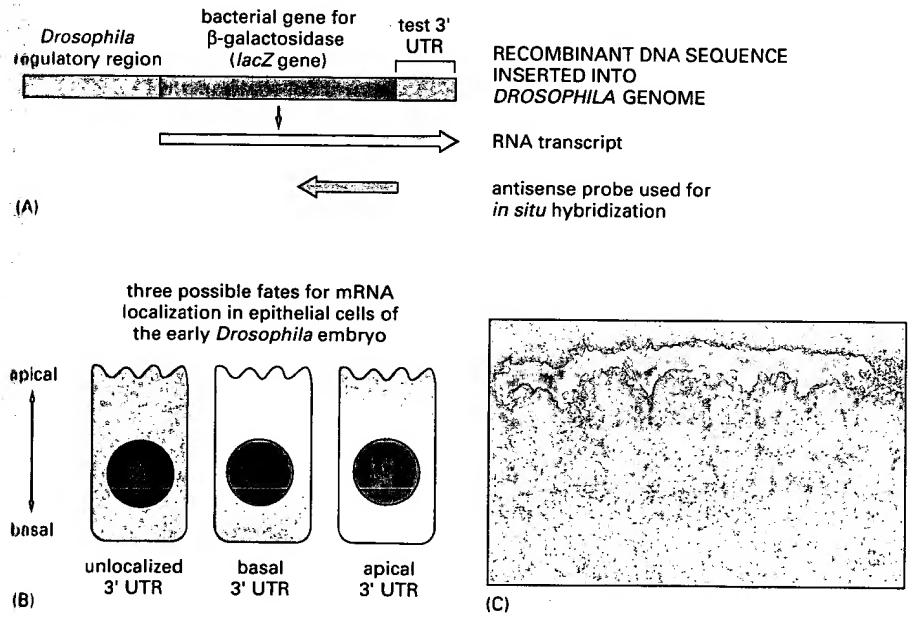


Figure 9–79 The importance of the UTR in localizing mRNAs to specific regions of the cytoplasm. *Drosophila* can be transfected with a recombinant DNA molecule coding for an mRNA in which a bacterial reporter sequence (encoding β-galactosidase) is linked to a chosen 3′ UTR sequence. According to the choice of 3′ UTR, the mRNA may be unlocalized in the embryonic cells, localized at their basal ends, or localized at their apical ends. (A) The recombinant DNA molecule used to test the effects of different 3′ UTRs. (B) *In situ* hybridization (for β-galactosidase RNA sequences) shows that the 3′ UTR determines the localization of the mRNA in the embryonic cells. (C) Photograph of an apically localized mRNA detected by *in situ* hybridization. The cells containing this mRNA are arranged in stripes along the axis of the embryo. (C, courtesy of David Ish-Horowitz.)

RNA Editing Can Change the Meaning of the RNA Message [60]

The molecular mechanisms used by cells are a continual source of surprises. An example is the phenomenon of ***trans* RNA splicing,** which occurs in all transcripts in the trypanosomes (the parasitic protozoa responsible for sleeping sickness). All mRNAs in trypanosomes possess a common 5′ capped leader sequence that is transcribed separately and added to the 5′ ends of RNA transcripts by splicing of two initially unconnected RNA molecules. *Trans* splicing is also used to add a 5′ leader to several mRNAs in nematodes and to combine the separate RNA transcripts that form the coding sequence of some chloroplast and mitochondrial proteins in plant cells. Cutting and pasting between RNA transcripts could speed up the evolution of new proteins, and the few cases where exons are known to be joined in this way are suspected to be evolutionary remnants of a much more extensive process that dominated ancient cells.

Another startling discovery is the process of **RNA editing,** whereby the nucleotide sequences of RNA transcripts are altered. In this process, discovered in RNA transcripts that code for proteins in the mitochondria of trypanosomes, one or more U nucleotides are inserted (or, less frequently, removed) from selected regions of a transcript, causing major modifications in both the original reading frame and the sequence, thereby changing the meaning of the message. For some genes the editing is so extensive that over half of the nucleotides in the mature mRNA are U nucleotides that were inserted during the editing process. The information that specifies exactly how the initial RNA transcript is to be altered is contained in a set of 40- to 80-nucleotide-long RNA molecules that are separately transcribed. These so-called *guide RNAs* have a 5′ end that is complementary in sequence to one end of the region of the transcript to be edited; this is followed by a sequence that specifies the set of nucleotides to be inserted into the transcript and then a continuous run of U nucleotides. The editing mechanism is surprisingly complex, with the U nucleotides at the 3′ end of the guide RNA being transferred directly into the transcript, as illustrated in Figure 9–80.

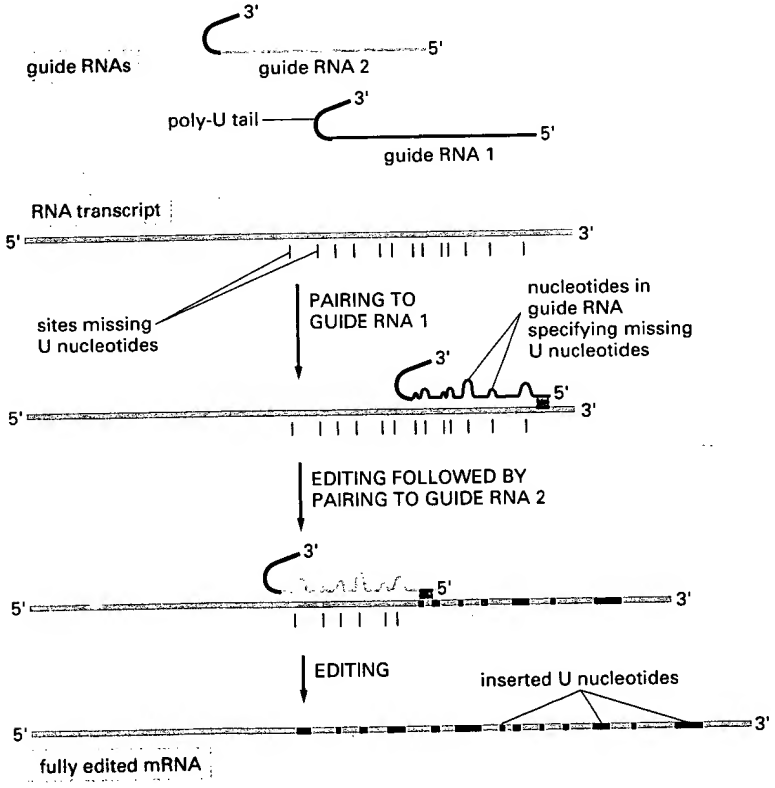


Figure 9–80 **RNA editing in the mitochondria of trypanosomes.** Guide RNAs contain at their 3′ end a stretch of poly U, which donates U nucleotides to sites on the RNA transcript that mispair with the guide RNA; thus the poly-U tail gets shorter as editing proceeds (not shown). Editing generally starts near the 3′ end and progresses toward the 5′ end of the RNA transcript, as shown because the "anchor sequence" at the 5′ end of most guide RNAs can pair only with edited sequences.

Extensive editing of mRNA sequences has also been found in the mitochondria of many plants, with nearly every mRNA being edited to some extent. In this case, however, bases are changed from C to U in the RNA, without nucleotide insertions or deletions. Often many of the Cs in an mRNA are affected by editing, changing 10% or more of the amino acids that the mRNA encodes.

We can only speculate as to why the mitochondria of trypanosomes and plants make use of such extensive RNA editing. The suggestions that seem most reasonable are based on the premise that mitochondria contain a primitive genetic system. There is evidence that editing is regulated to produce different mRNAs under different conditions, so that RNA editing can be viewed as a primitive way to change the expression of genes. Trypanosomes are extremely ancient single-celled eucaryotes, which diverged very early on from the lineage leading to plants, animals, and yeasts (see Figure 1–16). Perhaps, therefore, the extreme version of RNA editing found in their mitochondria is a holdover from very ancient cells, where most catalyses were carried out by RNA molecules rather than by proteins.

RNA editing of a much more limited kind occurs in mammals. The first case discovered involved the *apolipoprotein-B* gene, where RNA editing produces two types of transcripts: in one of these a DNA-encoded C is changed to a U, creating a stop codon that causes a truncated version of this large protein to be made in a tissue-specific manner. In another case a nucleotide change in the middle of an mRNA molecule changes a single amino acid in a transmitter-gated ion channel in the brain, significantly altering the channel's permeability to Ca^{2+}. For apolipoprotein-B the editing is catalyzed in a very straightforward way: a protein binds to a specific sequence in the mRNA and then catalyzes the deamination of a C to a U. It is not known whether the other cases of mammalian and plant RNA editing are protein-mediated in this way or whether, instead, they make use of short RNA templates, as in trypanosomes.

We now turn to controls that operate on the translation of mRNAs into proteins.

Procaryotic and Eucaryotic Cells Use Different Strategies to Specify the Translation Start Site on an mRNA Molecule [61]

In bacterial mRNAs a conserved stretch of six nucleotides, the *Shine-Dalgarno sequence,* is always found a few nucleotides upstream of the initiating AUG codon. This sequence forms base pairs with the 16S RNA in the small ribosomal subunit and thereby correctly positions the initiating AUG codon in the ribosome. This interaction makes a major contribution to the efficiency of initiation and provides the bacterial cell with a simple way to regulate protein synthesis. Many translational control mechanisms in procaryotes involve blocking the Shine-Dalgarno sequence, either by covering it with a bound protein or by incorporating it into a base-paired region in the mRNA molecule.

Eucaryotic mRNAs do not contain a Shine-Dalgarno sequence. Instead, the selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5′ end of the mRNA molecule, which is the site at which the small ribosomal subunit binds to the mRNA and begins scanning for an initiating AUG codon (discussed in Chapter 6). The nucleotides immediately surrounding the start site in eucaryotic mRNAs also influence the efficiency of AUG recognition during the scanning process. If this recognition site is poor enough, scanning ribosomal subunits will ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. This phenomenon, known as "leaky scanning," is a strategy frequently used to produce two or more proteins, differing in their amino termini, from the same mRNA. It allows some genes to produce the same protein with and without a signal sequence attached at its amino terminus, for example, so that the protein is directed to two different compartments in the cell.

Another important difference between eucaryotic and procaryotic translation is that eucaryotic ribosomes dissociate rapidly from the mRNA when translation terminates. Thus reinitiation at an internal AUG codon after translation of a preceding open reading frame is much less efficient in eucaryotes than in procaryotes. Together, these differences—scanning from the 5′ cap and a limited ability to reinitiate at internal AUG codons—explain why the vast majority of eucaryotic mRNAs encode only a single protein and why the first AUG codon from the 5′ end is usually the functional start site for translation.

A few eucaryotic cell and viral mRNAs initiate translation by an alternative mechanism that involves internal initiation rather than scanning. These mRNAs contain complex nucleotide sequences, called *internal ribosome entry sites,* where ribosomes bind in a cap-independent fashion and start translation at the next AUG codon downstream. The details of this mechanism are not known.

The Phosphorylation of an Initiation Factor Regulates Protein Synthesis [62]

Eucaryotic cells decrease their overall rate of protein synthesis in response to a variety of situations, including the deprivation of growth factors, infection by viruses, heat shock, and entry into M phase of the cell cycle. Much of this regulation is thought to involve the initiation factor **eIF-2,** which is phosphorylated by specific protein kinases to decrease the overall rate of protein synthesis.

The normal function of eIF-2 is outlined in Figure 9–81. This protein forms a complex with GTP and mediates the binding of the methionyl initiator tRNA to the small ribosomal subunit, which then binds to the 5′ cap of the mRNA and begins scanning along the mRNA. After an AUG codon is recognized, the bound GTP is hydrolyzed to GDP by the eIF-2 protein, causing a conformational change in the protein and releasing it from the small ribosomal subunit. The large ribosomal subunit then joins the small one to form a complete ribosome that begins protein synthesis.

Because eIF-2 binds very tightly to GDP, a guanine nucleotide releasing protein (see Figure 15–50), designated eIF-2B, is required to cause GDP release so that a new GTP molecule can bind and eIF-2 can be reused (Figure 9–82A). The reuse of eIF-2 is inhibited when it is phosphorylated because phosphorylated eIF-2 binds to eIF-2B unusually tightly, preventing the completion of nucleotide exchange. There is more eIF-2 than eIF-2B in cells, and even a fraction of phosphorylated eIF-2 can trap nearly all of the available eIF-2B, thereby preventing the reuse of even the nonphosphorylated eIF-2 and greatly slowing protein synthesis (Figure 9–82B).

When the activity of a general translation factor, such as eIF-2, is reduced by phosphorylation, one might expect that the translation of all mRNAs would be reduced equally. Contrary to this expectation, however, the phosphorylation of

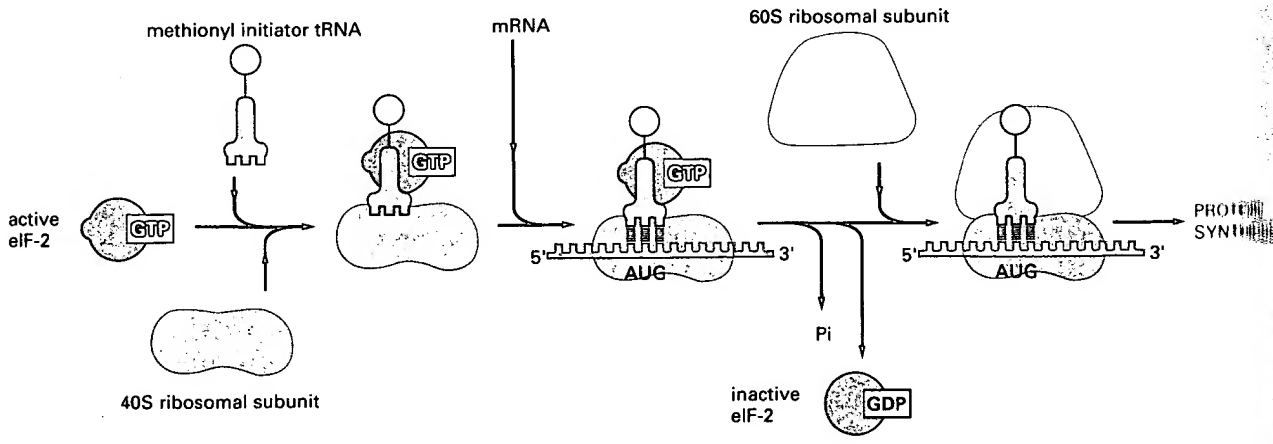


Figure 9–81 **The role of eIF-2 in the initiation of protein synthesis.**

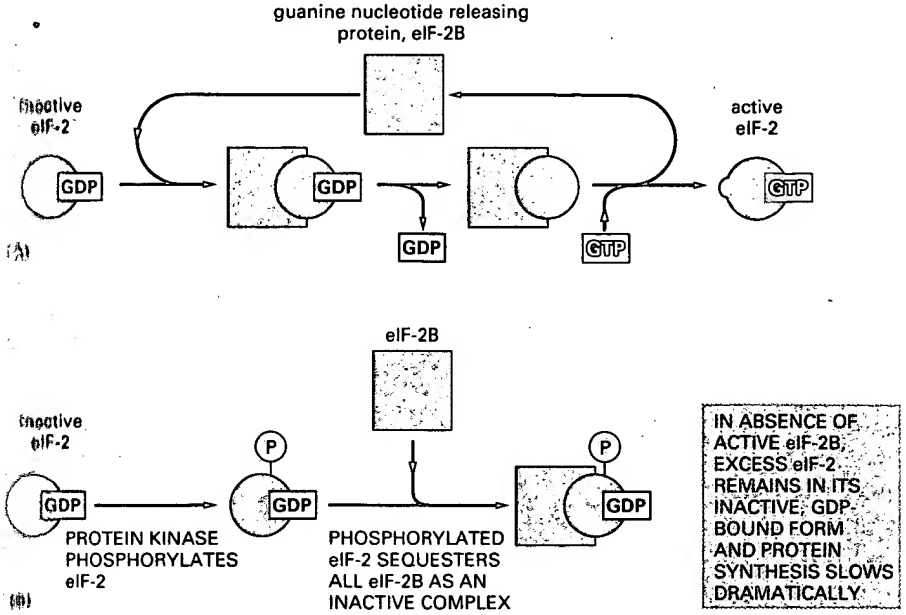


Figure 9–82 The eIF-2 cycle. (A) The recycling of used eIF-2 by a guanine nucleotide releasing protein (eIF-2B). (B) eIF-2 phosphorylation controls protein synthesis rates by tying up eIF-2B.

IF-2 can have selective effects, even enhancing the translation of specific mRNAs. This can enable yeast cells, for example, to adapt to starvation for specific nutrients by shutting down the synthesis of all proteins except those that are required for synthesis of the missing nutrients. The details have been worked out for a specific yeast mRNA that encodes a protein called GCN4, a gene regulatory protein that is required for the activation of many genes encoding proteins that are important for amino acid synthesis. The GCN4 protein is produced by a specific activation of the translation of its mRNA following amino acid starvation that is induced when eIF-2 becomes phosphorylated. By a complex mechanism depending on competition between correct and incorrect ("decoy") sites of initiation of translation near the 5′ end of the GCN4 mRNA, the reduction of eIF-2 activity actually leads to an increase in the synthesis of the GCN4 protein.

Regulation of the level of eIF-2 is also important in mammalian cells as part of the mechanism by which they can be induced to enter a nonproliferating, resting state (called G_0) in which the rate of protein synthesis is reduced to about one-fifth the rate in proliferating cells (discussed in Chapter 17).

Proteins That Bind to the 5′ Leader Region of mRNAs Mediate Negative Translational Control [63]

The translation of some mRNA molecules is blocked by specific *translation repressor proteins* that bind near the 5′ end of the mRNAs, where translation would otherwise begin. This type of mechanism is called **negative translation control** (Figure 9–83). It was first discovered in bacteria, where it enables excess ribosomal proteins to repress the translation of their own mRNAs—a form of negative feedback regulation.

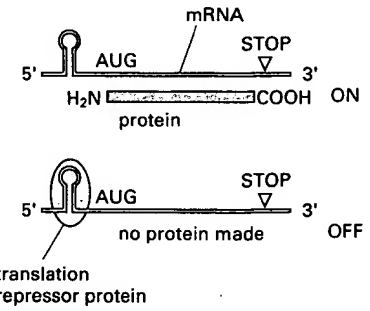
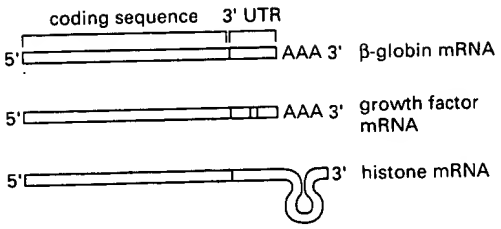


Figure 9–83 Negative translational control. This form of control is mediated by a sequence-specific RNA-binding protein that acts as a translation repressor. Binding of the protein to an mRNA molecule decreases the translation of the mRNA. Several cases of this type of translational control are known; the illustration is modeled on the mechanism that causes more ferritin (an iron storage protein) to be synthesized when the free iron concentration in the cytosol rises; the iron-sensitive translation repressor protein is called aconitase (see also Figure 9–86).



	HALF-LIFE
β-globin mRNA	> 10 hours
growth factor mRNA	30 minutes
histone mRNA	1 hour when cell is synthesizing DNA, but 12 minutes when cell is not synthesizing DNA

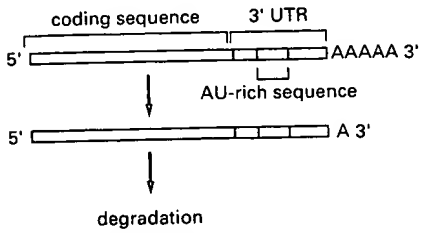
RNA stability. Three normal mRNAs with very different half-lives. The continuous rapid degradation of the mRNA molecules that encode various growth factors allows their concentration to be changed rapidly in response to extracellular signals. A signal for degradation in the 3′ untranslated region (UTR) determines the half-life of the growth factor RNA (see Fig. 9–85). Histones are needed mainly to form the new chromatin produced during DNA synthesis; a large change in the stability of their mRNAs helps to confine histone synthesis to the S phase of the cell cycle.

In eucaryotic cells a particularly well-studied form of negative translational control allows the synthesis of the intracellular iron storage protein *ferritin* to be increased rapidly if the level of soluble iron atoms in the cytosol rises. The iron regulation depends on a sequence of about 30 nucleotides in the 5′ leader of the ferritin mRNA molecule. This *iron-response element* folds into a stem-loop structure that binds a translation repressor protein called *aconitase*, which blocks the translation of any RNA sequence downstream (see Figure 9–83). Aconitase is an iron-binding protein, and exposure of the cell to iron causes it to dissociate from the ferritin mRNA, releasing the block to translation and increasing the production of ferritin by as much as 100-fold.

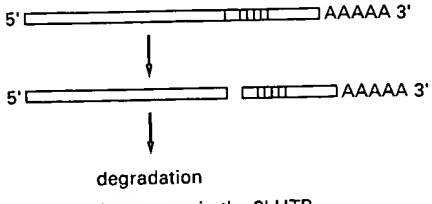
Gene Expression Can Be Controlled by a Change in mRNA Stability [64]

Most mRNAs in a bacterial cell are very unstable, having a half-life of about 3 minutes. Because bacterial mRNAs are both rapidly synthesized and rapidly degraded, a bacterium can adapt quickly to environmental changes.

The mRNAs in eucaryotic cells are more stable. Some, such as that encoding β-globin, have a half-life of more than 10 hours. Others, however, have a half-life of 30 minutes or less. The unstable mRNAs often code for regulatory proteins, such as growth factors and gene regulatory proteins, whose production levels change rapidly in cells (Figure 9–84). Many of these RNAs are unstable because they contain specific sequences that stimulate their degradation. A long sequence rich in A and U nucleotides in the 3′ untranslated region (UTR) of several mRNAs, for example, can, if transferred to other stable mRNAs by recombinant DNA techniques, cause them to be unstable. This AU-rich sequence appears to accelerate mRNA degradation by stimulating the removal of the poly-A tail found at the 3′ end of almost all eucaryotic mRNAs. Other unstable mRNAs contain recognition sites in their 3′ UTR for specific endonucleases that cleave the mRNA (Figure 9–85).



an evolutionarily conserved 50-nucleotide AU-rich sequence in the 3' UTR promotes the removal of the poly-A tail and causes the mRNA to become unstable



a repeated sequence in the 3' UTR promotes cleavage of the 3' UTR by a specific endonuclease. The fragments are rapidly degraded

Control of RNA degradation. Special sequences in the 3′ untranslated region (UTR) of unstable mRNAs are responsible for their unusually rapid degradation. As indicated, AU-rich sequences found in the 3′ UTR of many short lived mRNAs cause a rapid removal of the poly-A tail, which in turn makes the RNA unstable. Other mRNAs have sequences in their 3′ UTR that act as sites for specific endonuclease cleavage.

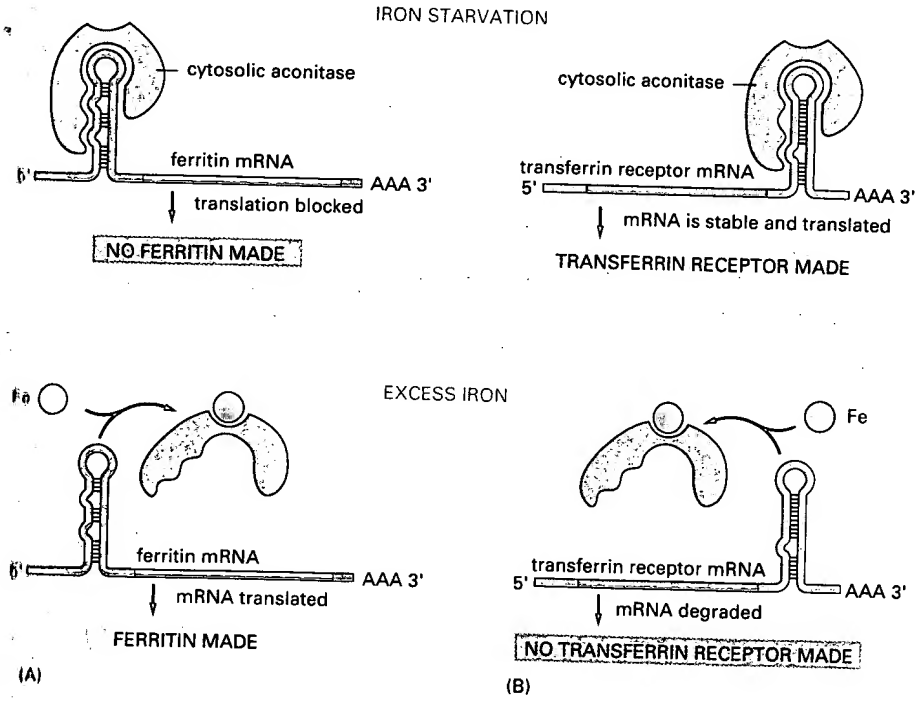


Figure 9–86 **Two posttranslational controls mediated by iron.** In response to an increase in iron concentration in the cytosol, a cell increases its synthesis of ferritin in order to bind the extra iron (A) and decreases its synthesis of transferrin receptors in order to import less iron (B). Both responses are mediated by the same iron-responsive regulatory protein, aconitase, which recognizes common features in a stem-and-loop structure in the mRNAs encoding ferritin and transferrin receptor. Aconitase dissociates from the mRNA when it binds iron. Because the transferrin receptor and ferritin are regulated by different types of mechanisms, their levels respond oppositely to iron concentrations even though they are regulated by the same iron-responsive regulatory protein. (Adapted from M.W. Hentze et al., *Science* 238:1570–1573, 1987; J.L. Casey et al., *Science* 240:924–928, 1988.)

The stability of an mRNA can be changed in response to extracellular signals. Steroid hormones, for example, affect a cell not only by increasing the transcription of specific genes, but also by increasing the stability of several of the mRNAs encoded by these genes. Conversely, the addition of iron to cells decreases the stability of the mRNA that encodes the receptor protein that binds the iron-transporting protein transferrin, causing less of this receptor to be made. Interestingly, the stability of the transferrin receptor mRNA seems to be modulated by the iron-sensitive RNA-binding protein aconitase, which, as we discussed above, also controls ferritin mRNA translation. Here aconitase binds to the 3′ UTR of the transferrin receptor mRNA and causes an increase in receptor production, presumably by inhibiting the function of sequences that otherwise cause rapid degradation of the mRNA. On the addition of iron, aconitase is released from the mRNA, decreasing mRNA stability (Figure 9–86).

Selective mRNA Degradation Is Coupled to Translation [65]

The control of mRNA stability in eucaryotic cells is best understood for the mRNAs that encode histones. These mRNAs have a half-life of about 1 hour during the DNA synthesis (S) phase of the cell cycle, when new histones are needed, but become unstable and are degraded within minutes when DNA synthesis stops. If DNA synthesis during S phase is inhibited with a drug, histone mRNAs immediately become unstable, perhaps because the accumulation of free histones in the absence of new DNA for them to bind increases the degradation rate of their mRNAs.

The regulation of histone mRNA stability depends on a short 3′ stem-and-loop structure that replaces the poly-A tail present at the 3′ end of other mRNAs (see Figure 9–84). A special cleavage reaction, which requires base-pairing to a small RNA in a ribonucleoprotein particle, creates this 3′ end after the histone mRNA is synthesized by RNA polymerase II. If the 3′ end is transferred to other mRNAs by recombinant DNA methods, they also become unstable when DNA synthesis stops. Thus, as for other types of mRNAs, the degradation rate of histone mRNA is strongly influenced by signals near the 3′ end, where mRNA degradation is thought to begin.